DOCUMENT RESUME

ED 337 475 TM 017 289

TITLE	Proceedings of the 1985 IPMAAC Conference on Public Personnel Assessment (9th, New Orleans, Louisiana, June 16-20, 1985).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	Jun 85
NOTE	222p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Assessment Centers (Personnel); *Evaluation Methods; Futures (of Society); Item Bias; Job Analysis; *Job Performance; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; *Public Sector; Screening Tests; Test Use
IDENTIFIERS	International Personnel Management Association

ABSTRACT

The International Personnel Management Association Assessment Council (IPMAAC) is a section of the International Personnel Management Association for individuals engaged in professional level public personnel assessment. Author-generated summaries/outlines of papers presented at the IPMAAC's 1985 conference are provided. Three papers are summarized in the presidential forum under the title "Future Perspectives". The keynote address is "Comparable Worth in Perspective" by T. A. Mahoney. Paper session titles include: "Report on the IPMAAC Job Analysis Project"; "Problems and Payoffs in Automated Applicant Tracking"; "The Role of Implementation in Personnel Management--Connecting Theory to Practice"; "Comparable Worth"; "Sex and Occupational Differences on the Perceived Importance of Wage and Salary Determinants"; "Practical and Theoretical Applications of Item Bias Studies"; "Automated Test Generation"; "Validity Generalization Summary"; "A Simplification of the Assessment Center Process through the Use of the Word Processor"; "Use of Ratings and Self Assessment in Selection"; "Departmental Ratings for Promotional Examinations"; and "A Systematic Approach to Determining Critical Job Behaviors". Symposia titles include: "Microcomputing in Personnel"; "Change Implementation Techniques for Public Institutions"; "The Use and Misuse of Item Bias Statistics"; "Biodata as an Alternative Selection Technique: An Extensive Evaluation"; "Putting Validity Generalization and Transportability to Optimal Use"; "Multipurpose Job Analysis"; "Validation, Implementation, Transportability, and Utility of a Selection Procedure for Professional Classes in a State Merit System"; and "The Use of Employment Selection Procedures with Large Multi-Ethnic and Racial Candidate Populations: Perspectives and Strategies". A student paper ("The Influence of Sex Stereotyping and the Sex of the Job Evaluator on Job Evaluation Ratings"); a special session ("Equitable Compensation: Methodological Criteria for Comparable Worth"); the Western Region Intergovernmental Personnel Assessment Council's "Historical and Future Perspectives on Assessment Centers"; the Personnel Testing Council's "Uniform Guidelines on Employee Selection Procedures: A Proposed Alternative"; a video presentation ("Use of Video in Assessment"); and an IPMAAC Professional Affairs Committee Forum ("Professional Ethics: Requirements, Issues, and Practicalities") are reviewed. An author index is provided. (SLD)

IPMA Assessment Council

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARIANNE ERNESTO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED337475

PROCEEDINGS OF THE

1985 IPMAC CONFERENCE

ON

PUBLIC PERSONNEL ASSESSMENT

JUNE 16-20, 1985

NEW ORLEANS, LOUISIANA

Published and distributed by the International Personnel Management Association (IPMA). Refer any questions to the Director of Assessment Services, IPMA, 1617 Duke Street, Alexandria, Virginia 22314, 703/549-7100.

PROCEEDINGS OF THE 1985 IPMA ASSESSMENT COUNCIL
CONFERENCE ON PUBLIC PERSONNEL MANAGEMENT

The PROCEEDINGS are published as a public service to encourage communication among assessment professionals about matters of mutual concern.

The PROCEEDINGS essentially summarize the presentations from information available to the Publications Committee of IPMAAC. Some presenters furnished papers which generally included extensions of their remarks, while others merely furnished a topical outline of their presentations. Adequacy and detail of information available varied greatly. For a few sessions no information was available from which a summary could be prepared.

Every attempt has been made to accurately represent each presentation. The PROCEEDINGS are summaries and condensations made by the reviewer(s). Persons wishing to quote results should consult directly with the author(s). In many cases, tables and statistical data and other supporting material were available which had to be excluded because of length. However, bibliographies are included if they were available.

PREPARED UNDER THE GENERAL DIRECTION OF:

Clyde J. Lindley
Associate Director, Center for Psychological Service
Chair, Publications Committee, IPMAAC

ASSISTED BY:

Thelma Hunt
Professor Emeritus of Psychology
George Washington University

Acknowledgment is also made of the work of Jamie Smith, graduate student at George Washington University.

IPMA ASSESSMENT COUNCIL

The INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL (IPMAAC) is a professional section of the International Personnel Management Association--United States for individuals actively engaged in or contributing to professional level public personnel assessment.

IPMAAC was formed in October 1976 to provide an organization that would fully meet the unique needs of public sector assessment professionals by:

- providing opportunities for professional development;
- defining appropriate assessment standards and methodology;
- increasing the involvement of assessment specialists in determining professional standards and practices;
- improving practices to assure equal employment opportunity;
- assisting with the many legal challenges confronting assessment professionals; and
- coordinating assessment improvement efforts.

IPMAAC OBJECTIVES support the general objectives of the International Personnel Management Association--United States. IPMAAC encourages and gives direction to public personnel assessment; improves efforts in fields such as, but not limited to, selection, performance evaluation, training, and organization effectiveness; defines professional standards for public personnel assessment; and represents public policy relating to public personnel assessment practices.

IPMAAC EXECUTIVE COMMITTEE
Bruce W. Davey, President
Susan K. Christopher, President-Elect
Doris M. Maye, Past President

Published and distributed by the International Personnel Management Association:

1617 Duke Street
Alexandria, Virginia 22314

(703) 549-7100

Refer any questions to Sandra Shoun, Director of Assessment Services.

TABLE OF CONTENTS

	Page
PRESIDENTIAL FORUM - Future Perspectives	1
Future Issues in Personnel	5
Comments and a Perspective on the Future of IPMAAC and the Work of Assessment Professionals	7
Future Directions of IPMAAC	10
KEYNOTE ADDRESS - Comparable Worth in Perspective	12
PAPER SESSION - Report on the IPMAAC Job Analysis Project	23
SYMPOSIUM - Microcomputing in Personnel	25
Introducing Computer Applications into Organizations	25
PAPER SESSION - Problems and Payoffs in Automated Applicant Tracking	29
Microcomputers for Conferences and Networking	33
SYMPOSIUM - Change Implementation Techniques for Public Institutions	36
The Use of "Stakeholders" in the Development of a Selection Procedure	36
PAPER SESSION - The Role of Implementation in Personnel Management - Connecting Theory to Practice	39
A Study of MMPI Use in Police Officer Screening	42
IPMAAC STUDENT PAPER AWARD - The Influence of Sex Stereotyping and the Sex of the Job Evaluator on Job Evaluation Ratings	48
IPMAAC SPECIAL SESSION - Equitable Compensation: Methodological Criteria for Comparable Worth	53
Technical Standards for Comparable Worth Implementation	53
PAPER SESSION COMPARABLE WORTH - The New Frontier for Public Human Resource Management	60
PAPER SESSION - Sex and Occupational Differences on the Perceived Importance of Wage and Salary Determinants	67
Psychometric and Selection Issues	71
The Myth of Proportional Representation	71
Further Support for Validity Generalization: A Test Publisher's Meta-analysis	72
An International Perspective of Personnel Selection Systems: British vs. American	77
SYMPOSIUM - The Use and Misuse of Item Bias Statistics	80
Difficulties with Delta	80

Page

PAPER SESSION - Practical and Theoretical Applications of Item Bias Studies 83
Application of a Latent Trait Approach to Detecting Item Bias 86
Item Bias Detection Methods for Small Samples 92
Physical Test for Firefighters 95

INVITED SPEAKER - Western Region Intergovernmental Personnel Assessment Council (WRIPAC)
Historical and Future Perspectives on Assessment Centers ... 96

SYMPOSIUM - Biodata as an Alternative Selection Technique: An Extensive Evaluation 101
Development, Validation, and Use of a State Police Entrance Exam in a Consent Decree Environment 106
Career Development Assessment Centers in Public Agencies 111

INVITED SPEAKER - Personnel Testing Council:
Uniform Guidelines on Employee Selection Procedures: A Proposed Alternative 114

PAPER SESSION - Automated Test Generation 120

SYMPOSIUM - Putting Validity Generalization and Transportability to Optimal Use 129
Application of Validity Generalization Within the United States Employment Service 129

PAPER SESSION - Validity Generalization Summary 134

SYMPOSIUM - Multipurpose Job Analysis 139
Multipurpose Job Analysis Works, But 139
Further Research on Assessment Centers 147
Selection of a Local City Official through an Assessment Center 147

PAPER SESSION - A Simplification of the Assessment Center Process through the Use of the Word Processor 154
Replicating Research on Police Promotional Assessment Centers 160

SYMPOSIUM - Validation, Implementation, Transportability and Utility of a Selection Procedure for Professional Classes in a State Merit System 167
Validation of the Professional Entrance Test (PET) for the State of Louisiana 167

VIDEO PRESENTATION - Use of Video in Assessment 172
Orientation to Assessment Centers - A Video Approach 172

IPMAAC PROFESSIONAL AFFAIRS COMMITTEE FORUM 173
Professional Ethics: Requirements, Issues and Practicalities 173

Page

PAPER SESSION - Use of Ratings and Self Assessment in Selection . 176
Supplemental Application Validation Based on Self-Rating and the Suitability of Subject Matter Experts and Raters. 176
The Validity of Self-Assessments Within a Police Sergeant Promotional System ... 180

PAPER SESSION - Departmental Ratings for Promotional Examinations ... 187

SYMPOSIUM - The Use of Employment Selection Procedures with Large Multi-ethnic and Racial Candidate Populations: Perspectives and Strategies ... 192
A Methodology to Determine Job Required Reading Levels ... 192

PAPER SESSION - A Systematic Approach to Determining Critical Job Behaviors ... 194
Strategies and Outcomes of Developing a Written Examination for a Large Multi-ethnic and Multi-racial Candidate Population ... 195
Defending Selection Decisions ... 198
Union Challenges to the Use and Interpretation of Promotional Examinations ... 198
Resolving Affirmative Action and Assessment Conflicts: One Jurisdiction's Journey Through the Realm of the Possible ... 203

AUTHOR INDEX ... 213

PRESIDENTIAL FORUM

Future Perspectives

Doris M. Maye, IPMAAC President, State Merit System of Personnel Administration, Atlanta, Georgia

Trends Affecting The Profession

Professional trend watchers, most notably John Naisbitt in Megatrends, point to national shifts toward decentralization, "multiple-options", and high technology in an environment characterized by accelerated movement from an industrial to an information society.[1] Education is becoming a life long "self-learning" activity that is moving from the formal institution into the mainstream of society, especially the work-place; and it is predicted that local and state governments will gain in power as Federal influences decrease.[2] EEO legislation will still be the "law of the land" and selection will continue receiving scrutiny, but the national attention of the '70's on racial discrimination is waning, relatively speaking, and the issues most likely to be litigated in the '80's include disparate impact, "comparable worth" claims and individual claims of sexual harassment, discrimination in performance appraisals and age discrimination.[3] Conflicting views will be held regarding the value of traditional assessment specialists. On the one hand, as more and more employers feel the sting of court action, one would expect to see increasing realization that "preventive maintenance" is the most prudent course of action--that it is significantly less costly to have qualified persons review, cure and maintain nondiscriminatory personnel practices than to pay an attorney to defend the employer.[4] On the other hand, as resources decline organizations appear to single out traditional assessment, especially selection, for retrenchment above and beyond across-the-board cuts in human resources because of a perceived decreased need because of decreasing hiring.[5] There appears to be a limited understanding that, in reality, it is necessary that assessment professionals actually do a better job since, "during a time of declining resources, selection mistakes are more visible, less tolerable, and less likely to be corrected by self-selection."[6] There also continues to be refinement in data pertaining to what contributes to an effective, well-run organization. Increasing emphasis is upon the interactive effects of the various process components and the value of a comprehensive "systems" approach to improving organizational functioning. For example, Jim Springer in the IPMAAC "Special Issue" of Public Personnel Management recently discussed the "human infrastructure" of organizations and argued for the concept of "total system validity".[7] The argument is paralled by many, but few so forcefully as Sheldon Zedeck and Wayne Cascio in the latest Annual Review of Psychology ... "consideration of selection issues is nonpragmatic and theoretically bankrupt unless consideration is also given to issues such as organizational design, motivation, career pathing, training and the like".[8] Professional trends point to broad, coordinated programs within the human resource function and between the human resource function and other operating programs.

Concurrently, there is an increasing emphasis on standards. Economic conditions and national trade deficts make "productivity" the watchword of the '80's. The assessment specialist is going to be called upon to focus his/her expertise on a wider and wider range of problems, at the same time that the internal standards of the assessment profession become more and more refined. Witness the developmental controversy over the Joint Technical Standards. Early draft versions of these revised professional testing standards were academically-oriented, relatively unsympathetic to practical application and economic constraints, and were focused at a sophisticated level of technical expertise. While the final version eases the constraints somewhat, the challenge to the assessment professional to refine and increase technical competencies is very much a reality. High technology will find more and more applications in assessment and will become more and more indispensible to the broadened functions. Even now, few organizations can operate a human resource program without computerization; and complex computerization is inherent in adaptive testing and other "new wave" methodologies. Measurement theories and practices existing essentially unchanged for decades are being superceded by concepts, many of which have existed for some time, now made more operationally feasible for broader applications through high technology. The assessment specialists will likely find it a difficult task to forestall technical obsolescence, to balance increased demands with decreased resources, and to reconcile technical and social demands -- all in a "new" information-oriented world.

IPMAAC's Role

How then does the assessment professional proceed? The Number One strategy is KNOWLEDGE -- Knowledge of what the big picture is (national realities/trends, organizational realities/trends, professional realities/trends); knowledge of how his/her organization fits into the big picture (what is the purpose of the organization, what forces are affecting the organization's direction and what are the methods of accomplishing the organization's purpose); knowledge of human resource systems and the applications of assessment in both the system components and in integrating the system (what are the prerequisite technical skills, how/where does one obtain these skills) and knowledge of how to operationally apply skills in on-going activities (what procedures and instruments exist, how to assess appropriateness for use, and how to demonstrate the value of these skills and methodologies to the organization). To my mind, this context defines IPMAAC's role.

As a professional organization IPMAAC is primarily a resource -- valuable as a conduit for knowledge, but even more valuable as a focused contributor to the body of knowledge. This resource could be narrowly or broadly defined. The advantages of a narrow focus would be in-depth probing of limited topics with the potential development of highly refined skills and methodologies. For example, the largest identified group of IPMAAC members comes from local public jurisdictions. Likewise, the largest identified group of potential IPMAAC members is likely to also be local public jurisdictions. Local public jurisdictions are gravely concerned about selection and promotion of police and fire personnel. It is a costly, highly litigious area and the more information

available about how to hire and advance in these areas, the better it is for the local jurisdictions and for the jurisdictions' professional staff in the specifically defined technical roles.

But is it most desirable that IPMAAC be defined as the association of police and fire selection specialists? I think not. Rather, the focus should remain "Assessment" -- specifically "Applied Assessment", and strategies should be brought to bear to more consistently evolve the organization toward the broader perspective envisioned by its founders. From this perspective I refer to the IPMAAC Bylaws:

> "the purpose of this Section shall be ... to encourage ... and give direction to ... assessment ... efforts in such fields as, but not limited to, selection, performance evaluation, job analysis and organizational effectiveness."[9]

The binding commonality is the wide range of assessment activities pertaining to theories and practices of the people/work interaction, its measurement and intervention.

Additionally, the operating institutional influences on assessment should be the lesser of several spheres of emphasis. The profession cannot afford to be overly exclusive. There is too much to do and too few resources available to the task to allow the situationally specific factors that certainly warrant recognition and attention to become focus-defining. Virginia Boehm, in her address to IPMAAC during the Sixth Annual Conference, concluded that there is no substantive difference between public and private sector assessment. Further, she urged a pooling of resources to deal with fundamental professional issues and problems.

> "The challenge of moving beyond a narrow task focus to a broader view of work and consequently also moving beyond ability and proficiency to a broader view of work-related assessments cuts across employment sectors ... It is only through unified effort that the challenges facing assessment can be successfully met. Competing social priorities, budget cuts and recessions, and the need to reconceptualize our view of work and of assessment are enormous challenges that require all the expertise that all of us possess."[10]

I say "Amen" -- and let us clearly focus on the goal so that our interim steps can be straight, sure and toward the mark.

Reference Notes

1 Naisbitt, John, Megatrends: Ten New Directions Transforming Our Lives. New York: Waiver Books, 1982.

2 Ibid.

3 Sutter, Lloyd, Employment Discrimination Law. Paper presented to the Metropolitan Atlanta Chapter of the International Personnel Management Association, Atlanta, March 1983.

4 Ibid.

5 Boehm, Virginia, Public and Private Sector Assessment: Is there a Difference? Paper presented at the annual conference of the International Personnel Management Association Assessment Council, Minneapolis, June 1982.

6 Ibid.

7 Springer, James, A Bridge Collapse and Personnel Selection, Public Personnel Management, Volume 13, Number 4, Winter 1984.

8 Zedeck, Sheldon and Wayne Cascio, Psychological Issues in Personnel Decisions, Annual Review of Psychology, Volume 35, 1984, page 463.

9 Bylaws of The International Personnel Management Association Assessment Council, page 1.

10 Boehm, Ibid.

* * *

Future Issues In Personnel

Barbara Showers, IPMAAC Past-President, Department of Regulation and Licensing, State of Wisconsin

I see myself more as a facilitator than as a visionary, however, when pressed into thinking about the future in our profession, I began to see an apparent trend in current political and social policy, which would be healthy to us as a profession to pursue whether it remains a trend or not.

I am referring to the current trend in government and politics toward managerial responsiveness, and away from the imposition of complex legal requirements and guidelines; toward private entrepreneurial values and away from bureaucratic governmental values.

For examples of this trend, consider the reduction of the federal Merit System Standards from a multipage document of specifics to a one page statement of general concepts which leave responsibility for carrying them out to the judgement of the state and local governments.

Or consider the impending study of the Uniform Guidelines. The primary question appears to be, "is all this specifically and requirement language necessary?"

These examples appear to value the concept of "let the managers manage."

If this currently is a trend, then we may have to readjust our values and priorities. By this I mean that it would be healthy for us now to expand our thinking in terms of how we can be most valuable to the management of our agencies, and not perhaps, how we must be the watchdogs of complex guidelines and procedures to keep management in check.

I am not suggesting that we abandon guidelines or professional standards. On the contrary, this may be an excellent opportunity to develop creative ideas and implement strong professional standards. I am suggesting that we won't be able to rely on the negative motivation of the need to satisfy guidelines, laws and rules and stay out of court. I have always thought this was an unhealthy attitude in government personnel agencies, though it has been a necessity at times.

In my experience, occupational testing has never been in the mainstream of management priorities. It is expected to work quietly and effectively in the background, and not impede the progress of the agencies' major objectives. The only attention we get is when something goes wrong, like a major suit. From the reading I have done, this is apparently true of personnel in general, not just selection. Although if personnel

in general is undervalued by managers, then selection must be particularly uninteresting to most of them.

While we may find comfort in being so easily out of the mainstream, it is not a healthy place to be today. If the managers are going to be directed to manage, then we must have the influence to assure that the ethical standards of testing and merit selection which we hold and value are valued by our managers. To value our standards, the managers must value us. They must see us as a value to them and their objectives.

How can we increase our value to management? Some may say these ideas have always been around, but I think they are just now beginning to receive the attention they deserve, at this conference, and in the literature. I will use management terms to describe these, then interpret them in terms of selection. The five are: efficiency, flexibility, innovation, communication, and services.

1. Efficiency: Improving selection efficiency through shortening selection processes, and increasing applicant processing and test development efficiencies via computer technology.

2. Flexibility: Increasing hiring flexibility through broader certification rules, including group certification, categories, and expanded certification for affirmative action.

3. Innovation: Improving the quality and attractiveness of our tests via innovative testing approaches such as computer adaptive testing and simulation testing, and assessment centers.

4. Communication: Talking about our tests in terms managers understand and value--utility, productivity improvements, quality of hirers.

5. Services: Expanding services to management in related skill areas such as management surveys, comparable worth, and performance evaluation. The somewhat light turnout on some of the comparable worth sessions indicate we may not think of this as our area. But job evaluation systems are similar technically to what we do, and we could be useful in these areas.

It seems clear that the pendulum is swinging away from an emphasis on mechanisms and legalities in government toward an emphasis on managerial efficiency, effectiveness, and responsiveness. Away from the values of bureaucracy toward the values of private enterprise. Though most of us work in government settings, we could do well to seek out and learn from our colleagues in private enterprise. They may have a lot to tell us, both good and bad. A scan of our conference program topics and the special IPMAAC issue of the Public Personnel Management indicates that many already are making strides in the areas I have identified, e.g., more comparable worth, performance evaluation, and training applications in professional selection.

Over the past decade, we have educated our managers and lawyers to the technical terms of test validation, cross validation, generalizability, transportability, and coefficients of correlation. Now we must educate ourselves to the management concepts of efficiency, flexibility, innovation, communication, and services. These are not mutually exclusive. I see us enhancing our professional values by cultivating management values and becoming part of the management team, rather than choosing between the two in order to retain "professional purity."

* * *

PRESIDENTIAL FORUM (continued)

Comments And A Perspective On The Future Of IPMAAC And The Work Of Assessment Professionals.

Charles F. Sproule, IPMAAC Past-President, State Civil Service Commission, Harrisburg, Pennsylvania

As part of the Presidents' Forum, I have been asked to comment and provide a perspective on the future for:

1. IPMAAC, and
2. Our work as assessment professionals.

I also plan to comment on this question: What do we as an organization and as individuals need to focus on and do?

Future of IPMAAC

My career in assessment began in the early '60's. At that time most public sector assessment professionals were very isolated from their counterparts in other public jurisdictions. The U.S. Department of Health, Education and Welfare, and later the International Personnel Management Association, and Consortia stimulated a movement towards cooperation and sharing in the field of assessment. It was that movement which led to the establishment of IPMAAC.

Following are three recommendations on the future focus of IPMAAC:

1. Continue to serve as a forum and organization for assessment professionals to share information and resources. Expand and improve this type of service. Help us to learn from one another.

   Many successes relating to this objective have been achieved and good progress is evident from such efforts as the Winter 1984 special issue of Public Personnel Management on "Assessment Techniques and Challenges", the new "HACKER" and "PASS" newsletter

and the planned future monographs which we have learned about at this conference.

2. Provide a resource and leadership role in the education and training of assessment specialists.

IPMAAC has begun to achieve successes in this area such as: the recent seminar series on "Examination Planning", the completion of a national task survey of Assessment Specialists, the work being done on organizing and presenting a summary of information, research and materials on structured oral examining, and the planned future seminar series on ratings of training and experience.

Broader efforts should be considered for the future such as: a) Developing courses and seminars around a complete training plan for assessment professionals. Such a plan could be derived from the national job analysis which is underway to determine the performance requirments and KSA's needed. Such courses, seminars, and training aids could be developed by groups of jurisdictions or consortia under the guidance and sponsorship of IPMAAC and be offered nationwide. b) Proposing a graduate school curriculum for the practical education of future assessment professionals. c) Developing assessment tools which can help us identify our individual training needs.

3. Continue to actively work to influence laws, regulations and standards.

Again, IPMAAC has had successes in this area such as the existing IPMAAC standards for sharing test materials, and the improvement obtained in the Uniform Guidelines and improvements made in the recently published test standards as a result of IPMAAC involvement in evaluating drafts of these documents and recommending improvements.

There continues to be a need to represent practical and operational public sector assessment concerns. Some aspects of existing regulations and standards do not adequately address IPMAAC concerns. For example, an EEOC representative recently stated that the new test standards can be interpreted as supporting a toughening of the guidelines and requiring all three kinds of validity evidence (APA Monitor, May 1965, p. 20). Improvements are also needed in the definition of content validity and the standards for validity generalization.

Efforts to influence and improve regulations and standards tend to be time consuming and frustrating. However, this is an essential activity which we need to continue with vigor.

Personal Efforts

We, as individuals, as members of IPMAAC, and as employees of public sector jurisdictions or the private sector, need to also take actions to improve assessment and our profession. Some thoughts on this follow.

1. We need to think of our work and approach it with a resource allocation philosophy. That is, we need to work towards short-term practical and achievable objectives, as well as work on larger and more long-range issues.

   Examples of short-term efforts might include sharing our work products (e.g. job analysis reports, validation studies, new procedures or work methods, test items, rating scales, etc.) under security and exchange agreements, writing articles on areas of need or on new developments for PASS, sharing computer software via the HACKER, development of video-tapes for training examiners, development or refinement of a segment of an item bank or a selection process for a particular occupational area, etc.

   Examples of long-term efforts might include research on validity generalization or a particular measurement method (such as self-assessment ratings of training and experience), involvement in developing standards and training for assessment professionals, establishment of a multi-jurisdictional center for personnel assessment research, and development of taxonomies. We need to work as individuals through our employers and professional organizations to improve regulations, guidelines and standards.

2. We need to take a systems approach. That is we need to conceptualize assessment as an integral part of all of personnel management and not limit our perceptions and applications to selection. For example, our job analysis studies need to be tied to and be useful in developing training programs, setting pay rates, setting performance standards, and providing data for collective bargaining or personnel system corrections (e.g. class structure and class standards recommendations).

Perspectives Of The Future

President Doris Maye has described some trends which will influence our future. I agree with her perceptions and offer these additional reflections.

1. We can expect a continuation of the recent doing "more with less" philosophy of government. There will be a continued emphasis on production. This will provide a continuing challenge for us to improve and maintain quality while increasing the quantity of our products with less resources. This will challenge our innovation and creativity and cause us to re-evaluate how we accomplish our work.

2. The information and technology explosion will require us to continually re-educate ourselves. A major assessment innovation will be the development of integrated audio-visual test modes with computer/candidate interaction.

3. There will be continued debate on the adequacy and fairness of assessment techniques. However, this debate will be less in the forefront of personnel management concerns in the 80's and 90's.

4. There will be continued trends toward less rigid merit systems and more flexibility, more decision making authority and more accountability for line managers. This will lead to more variability in public sector merit systems. We need to be aware of this trend and build assessment systems which adhere to both merit concepts and provide flexibility. For example, we need to be more flexible in how test scores are used to calculate grades and consider alternatives to absolute ranking of candidates.

The trends and directions outlined above show that there is a great need for assessment professionals to be involved in and contribute to their professional organizations and what is happening in our field. I urge you to become active participants and bring new ideas, enthusiasm and creativity to our work. Do your fair share. Take time to communicate your successes and failures to your colleagues. We need you.

* * *

PRESIDENTAL FORUM (continued)

Future Directions of IPMAAC

Bruce W. Davey, IPMAAC President-Elect,
Connecticut State Personnel Department, Hartford

I want to first say that my personal opinion is that I really shouldn't be up here right now. I'm the new kid, and I haven't earned my battle scars and bruises yet like my colleagues here have.... Well, I've gotten a few already, but not very many. So I think I really should just be sitting someplace at a distance absorbing the collective wisdom of this session so I can get some guidance that might help me to live up to the tradition that's assembled here.

But I am up here and that being the case, I'll talk a little bit about future directions as I see them from my pre-battle-scar perspective.

In the past year the conitnuity committee conducted a membership survey to try and identify how IPMAAC's members felt about IPMAAC's products and services, and what you folks thought we should be doing in the

future. As I said in the opening session, we'll try to get those results into the Assessment Council News or distributed with the Assessment Council News, but anyway we'll get the full survey results to you. But for purposes of this forum, I'll highlight a little.

Not surprisingly, the survey indicated that IPMAAC should be - or should continue to be - a vehicle for communication and education. Our major charge is to keep the membership informed on key issues in assessment. This we primarily do through the Annual Conference and the Assessement Council News. We also have a few smaller publications and this year will be adding some monographs or articles on assessment topics.

The membership also wants a voice, an organization to speak for them on issues such as sound assessment methods, professional ethics and related issues. And here I don't really know whether IPMAAC should be doing anything different or not.

I think the quality of our members is such that we have a pluralistic voice. We don't have a single unified voice and I hope we never do. When we have one voice we'll have nothing new to say to one another. So I project us continuing in these directions and also some others. I see us strengthening our relationship with IPMA over the next few years--in part because it's a maturing relationship in which we're kind of learning how to deal with one another, and in part because I think that the role of assessment in personnel is growing; and, the natural outgrowth of that should be that the role of IPMAAC within IPMA is growing. Our discipline is spreading. We're even beginning to infiltrate the ranks of the classifiers. Job evaluation is coming under scrutiny because of the comparable worth movement. And there are lots of other potential points of "infiltration" or growth. Maybe performance evaluation will be the next new and sexy topic. Or maybe it will be productivity measurement. I don't know. But clearly we're going to grow.

* * *

KEYNOTE ADDRESS

Chair: William E. Tomes, South Carolina Division of Human Resource Management

Comparable Worth in Perspective

Thomas A. Mahoney, Owen School of Management, Vanderbilt University, Nashville, Tennessee

Comparable worth has been characterized as "the looniest idea since Looney Tunes hit the screen," as "the Civil Rights issues of the 80's," and as "a moral idea that deserves to be taken seriously." It is not very flattering to be asked to speak about a looney idea, but it is quite flattering to be asked to address a moral issue. No idea that generates so much concern and attention can be ignored. The concept of comparable worth arouses scorn, ridicule, fear, enthusiasm, and even a somewhat moral/religious fervor. Given this range of reactions, it can be neither ignored nor endorsed. But it does warrant considerable attention, particularly by people working in personnel and human resource management.

There is what I call a comparable worth movement and a comparable worth concept. The movement is evidenced by the social and political attention given to issues of woman's earnings, and the concept relates more directly to proposed definitions of discrimination in pay. Taken in context, I find the movement much easier to understand than the concept, and it is the concept that seems to be the source of much of the confusion about comparable worth and what it means. There will be later speakers who are more committed to positions on comparable worth than I am, so let me try to put the movement in perspective and to provide a background for understanding their positions.

I'll start first with a distinction between doctrine-theory-policy that I find useful in analysis of public policy proposals. Briefly, public policy must be viewed relative to certain problems or concerns of society; it is intended to correct and/or prevent specific problems. Doctrines are broad statements of belief and guides to behavior often motivated by ideological positions. And theory provides a basis for analysis of problems and the prediction of consequences of policy; it provides a test of whether or not policy proposals will impact upon the focal problems. The comparable worth concept or doctrine is being advanced as public policy--to evaluate it we must analyze the problems it addresses and theories of worth and wage determination. So much for the jargon, let's look at comparable worth.

Let's start with the slogans. The doctrine associated with comparable worth has been stated as "equal pay for jobs of comparable (read 'equal') worth." Now as doctrine, this principle has enjoyed wide acceptance for decades and even centuries, particularly if restated as "equal pay for equal work." The issue lies in the definition of "equal work" or "equal worth."

An early reference to pay in the Biblical parables evidenced this doctrine . . .

> Toilers in the vineyard who had worked all day objected when the employer paid those who worked only part of the day the same wages. The employer justified his action as his prerogative, and noted that all workers had, in effect, accepted employment at those terms.

A more recent account appeared in a newspaper column in the Spring of 1974 when comparison was made between the earnings of Morris the cat and American actors--

> According to informed sources, Morris, the finicky cat in that TV ad for the pet food, earned $10,000 last year, not to mention residuals.
>
> Which is about twice the income of the average American actor.
>
> Well, there you are. What makes money so fascinating a subject, after all, is the magnificent lack of justice with which it gets distributed.[1]

Since then, we have had the various concerns raised by comparable worth advocates and, more recently, the many critical examinations of the earnings provided to chief executives, examinations which question the equity of what are cased as outrageous and exorbitant earnings.

The point is that the doctrine of "equal pay for jobs of equal worth" seems to be relatively accepted by our society as a definition of equity. And that we constantly question what appear as inequities in relative compensation. In that sense, comparable worth is not a new issue, nor is it a gender issue. It is an issue of equity that has persisted over considerable time.

Equity is not a simple issue to address--it involves various subjective judgements.

> Note that equity is defined as "the absence of inequity"!

"Worth," "value," "equity" and "justice" have been the subjects of numerous philosophical disputes over time. One such examination was made by Adam Smith whom most of you identify as an economist. In fact, he was a teacher of moral philosophy! Smith identified two concepts of worth and value . . .

> Exchange value is what someone is willing to pay.
>
> Use value is value I impute to use.

---

[1] The Christian Science Monitor (June 20, 1974).

These two measures need not be the same and, in fact, won't be the same in any exchange. I buy a personal computer because its value to me is greater than the exchange value I have to pay, and I work at Vanderbilt because the exchange value for my services is greater than their use value to me. The concept of a market rate for exchange is a crude measure of value, but remember that use value and exchange value will vary considerably from one employer and one individual to another.

One of the common ways of addressing value for occupations and work has been through collective bargaining--the workers, through a union, agree that the relative pay for different jobs is equitable. If judged as inequitable, they negotiate for different rates.

Another common way of determining equitable relative wages is through job evaluation of some form which assigns relative wages to different jobs in an organization. Whether derived through negotiations or job evaluation, the real test of the resulting wage structure is whether or not the affected workforce accepts the wage structure. An employer who is unable to recruit and retain employees at the announced rates, or who is faced with constant grievances and slowdowns, will seek realignment of the wage structure such that it will be acceptable. Judgements about comparable worth are made every day in the evaluation of wage structures by employers, employees, and unions.

How then did the comparable worth movement ever develop as a gender issue? The concept of comparable worth is gender neutral, yet the movement is associated with women. To understand this, we must turn to an examination of the problem or issue motivating the comparable worth movement . . .

The comparable worth movement arose out of two related developments in our society during the period 1950 - 1980, the rapid introduction of women into the work force, and the civil rights movement.

As recently as 1950 the labor force was predominantly male, 72% were males. This was down from 82% in 1900.

Since 1950, women have gone to work outside the home. Now 60% of women work as compared with 29% in 1950, and women comprise 47% of the work force. Most of this increase came from married women, particularly women with children. Compared with 1950, the employment rate for women with children under six years increased from 12% to 45% and for women with children aged 6-19, it increased from 26% to over 60%.

This increased employment of women occurred as the job composition of the workforce was changing also. During this period, we had an expansion of employment in finance, trade, and particularly the services. Women found jobs in all of these industries, but notably in the service industries.

Forty-one percent of the female workforce is employed in the service industry--health care, education, real estate, and the like.

Nevertheless, women find, on average, that they earn only 60% of what men earn and this gender gap in earnings has occasioned concern. The gender gap suggests that women as a class do not have the economic power of men. Also, due to the increased divorce rate over the last 30 years, the proportion of children living with a single parent (usually female) rose from 10% in 1950 to 25% in 1980. And 50% of the children living with families with female heads are classified as living in poverty conditions.

Thus, despite the increased employment of women in the workforce, we find many of them heading families and living in poverty despite their new work roles.

Finally, when these working women read reports that lifetime earnings of a female college graduate are lower than lifetime earnings of a male high school dropout, it is not surprising that feelings of inequity are aroused.

Although not as clearly formulated as I have stated it here, comparable worth as a gender phenomenon is based upon social perceptions of inequity reflecting the observed gender gap in earnings. From a social standpoint, it can be argued that the objective of the comparable worth movement is reduction of the gender earnings gap between males and females. In this sense, comparable worth focuses upon gender earnings and not job or occupational worth. The comparable worth movement seeks a restructuring of incomes through elimination of a gender earnings gap.

LABOR FORCE COMPARISONS, 1950 AND 1980

	1950			1980		
	LABOR FORCE (%)	FEMALE LABOR FORCE (%)	PERCENT FEMALE	LABOR FORCE (%)	FEMALE LABOR FORCE (%)	PERCENT FEMALE
AGRICULTURE	20.2	4.5	6.3	9.9	2.7	11.7
MANUFACTURING	25.7	23.0	24.9	22.4	16.8	31.8
TRANSPORTATION, UTILITIES	10.3	7.4	20.2	7.3	4.2	24.7
TRADE	18.8	22.6	33.7	20.4	21.9	45.8
FINANCE, INSURANCE	3.4	5.0	40.7	4.3	8.2	58.0
SERVICE	18.0	34.2	53.0	28.7	41.1	61.1
PUBLIC ADMINISTRATION	4.4	4.1	26.2	5.3	5.1	40.8
	100.0	100.0	28.0	100.0	100.0	46.6

How then did comparable worth as a political movement ever come to address the relative worth of jobs and occupations? If the basic problem is one of gender earnings differentials, how will that be affected by realignments of occupational earnings? To understand this, let's digress to consider various theories and research into wage differentials and earnings differentials.

Theories of wage and earnings differentials generally tend to focus upon analysis of either labor supply characteristics or labor demand characteristics. Analysis of earnings differentials tend to look at differences in labor supply characteristics--some people earn more than others because they bring greater potential productivity to the job. This orientation tends to be called the human capital theory and argues that differences in education, training, and experience are related to differences in earnings. And empirical studies indicate that we can attribute much of the gender earnings gap to gender differences in human capital, depending upon how we define human capital. One set of analyses explained most of the lifetime earnings differential by measures of human capital including continuous labor force experience as a variable; other, more limited analyses explain about half of their earnings gap this way.

Human capital concepts have appeal to many, especially the more highly educated. Indeed, some associated with the comparable worth movement would elevate the human capital theory to a doctrine--people should be paid in accordance with their education. Many in the academic world sympathize with the doctrine--analyses of the worth of a PhD have suggested that it should be considered as consumption and not investment since it does not typically lead to higher lifetime earnings.

Other theories of wage differences focus upon characteristics of labor demand, particularly as evidenced in job or occupational demands. Certain jobs and occupations pay more than others because of their value to consumers and because of relative shortages in supply. A brain surgeon receives more than a cab driver because of consumer values and shortages of required skills. And a PhD in history driving a cab earns no more than a high school drop-out driving a cab. In short, these theories relate wage differences to differences in the valuation of the work performed rather than to the qualifications of the individual.

And this is how many would relate the gender earnings gap to concerns over the comparable worth of different occupations. Men and women tend to work in different occupations. For example, in a 1970 analysis of 553 occupations--310 were staffed with 80% or more males and 50 were staffed with 80% or more females. Conversely, 70% of males are employed in predominantly female occupations. And male occupations are paid more on the average than female occupations. For each 1% increase in the percentage of females in an occupation, average annual earnings declined $42 in 1970. And here lies the issue of the comparable worth phenomenon--women's occupations, it is argued, are undervalued relative to men's occupations.

The comparable worth movement seeks restructuring of gender earnings in society to raise women's earnings relative to men's earnings. And it seeks this outcome through restructuring occupational wages to increased wages of women's occupations.

Before looking at specific proposals, note one other thing not widely observed. In addition to occupational segregation of men and women, there also is industrial segregation. We noted earlier that women are concentrated in the service, financial, and retail industries. And there are historical industrial earnings differentials--auto and steel traditionally pay more than insurance and banking. A now classic analysis in the 1950s by John Dunlop illustrates the effect of industry upon wage rates. Dunlop examined the hourly wage rates paid to people in a single occupation--truck driver--in a single labor market--Boston--and found amazing differences, differences associated with the industry of employment. Delivery of certain products is valued more by consumers than delivery of other products and this is reflected in the relative worth of truck-driving jobs. A large part of the occupational segregation reflects an industry segregation as well--teachers, nurses, and clerks are found in the low paying industries and mechanics, jigsetters, and crane operators are found in the high paying industries.

One last element of background addresses the methods employed in seeking a restructuring of wage rates. Any restructuring of wage rates in the past typically occurred through collective bargaining or job evaluation. Either approach was employed when a sufficient proportion of an employer's workforce became upset enough to occasion problems of recruitment and/or turnover. The comparable worth movement achieved attention by seeking change through charges of wage discrimination under provisions of Title VII, a new tactic. Legal and political action have characterized the thrust of the comparable worth movement, not collective bargaining.

Employed women are not as well represented in collective bargaining as men. In part, because their industries and occupations aren't as well organized.

Legal action through charges of wage discrimination under Title VII is taken by women because this is a course open to them that has not been open to males. And one thrust of the comparable worth movement is to argue that the gender gap in earnings is a consequence of wage discrimination.

Wage discrimination has been recognized by the courts, but the extreme comparable worth principle has not been accepted yet. In fact, the court's refusal to rehear the Spaulding vs. Washington case hints that the court will not reject market wage as a measure of comparable worth.

Charges of wage discrimination employing comparable worth arguments call up issues of job evaluation, a process of wage determination which is not too well understood despite its widespread practice.

One charge made is that traditional job evaluation is biased against female-dominated occupations and should be changed in unspecified ways to correct for this bias and discrimination.

At the same time, others argue that some method of job evaluation is necessary to provide a true measure of comparable worth and eliminate dependence upon so called market rates which, it is argued, do not reflect true job worth.

Given the attention to job evaluation and the background and interests of this audience, let's digress to examine what job evaluation is and what can be expected of it.

In one sense, job evaluation is a way of determining use value to an employer, or relative use value of different jobs and occupations. There is nothing magical about it and the approach to valuation varies from one employer to another. The intent of job evaluation is to establish pay differentials which will be effective in the attraction of labor, the elimination of grievances, and concur with the evaluative judgements of the employer and the workforce. The only real test of validity of job evaluation lies in acceptance by those affected, and this is evidenced in modern approaches to job evaluation which employ policy-capturing analysis. The approach taken is to replicate the subjective judgements of the workforce and employer. In this sense, it is an alternative to bargaining over appropriate job differentials.

There are two different orientations to job evaluation which must be called out. One orientation is in the tradition of psychometrics. It approaches job evaluation much as one might approach the assessment of personality or intelligence. It begins with definition of the concept of job worth, the elaboration of dimensions of worth, and then the scaling to these dimensions. The major concern is one of establishing construct validity since an independent assessment of worth is believed lacking. Concerns for reliability, errors, and bias predominate.

The second orientation (which I share) is associated with institutional economics. This orientation assumes an independent criterion of worth and approaches job evaluation just as a selection oriented psychologist would approach validation of a test battery. Given a criterion, one searches for variables with predictive validity. And in that tradition, it is accepted that validity is time and situation bound--predictors must be revalidated whenever the situation changes.

Careful study of the history of job evaluation indicates that the logic in use applied was that of the institutional tradition. The most extreme example was the job evaluation plan of the steel industry in the 1940s and more recently the policy-capturing approaches of today. At the same time, job evaluation was rationalized with a reconstructed logic of the personality test developer. It was argued that dimensions of worth could be developed independent of an empirical criterion. Believing the reconstructed logic of job evaluation, many sought a

universal criterion of worth for job evaluation. In fact, it appears that the initial charge to the NAS committee assumed the existence of such a criterion.

Both approaches to job evaluation rest upon some form of value consensus, value consensus over the definition of worth in the psychometric orientation and value consensus over a occupational hierarchy in the institutionalist orientation. Since the consensus or lack of consensus over the occupational hierarchy is at the core of comparable worth concerns, consensus about the dimensions of worth is irrelevant unless an acceptable hierarchy is the outcome.

Note also that consensus of either form is most easily achieved in relatively smaller, more homogeneous populations. What is critical is consensus within an employer's workforce and not consensus throughout the American society. The relatively flat wage structure in the brewing industry and the highly differentiated wage structure of the auto industry evidence differences in values of different workforces and the impracticality of some form of national job evaluation.

Another thrust of the comparable worth movement has been more traditional in form. This has taken the form of employee pressure to restructure wage comparisons of individual employers, in particular public employers. A number of states have passed legislation directed toward establishing "pay equity" among jobs in public employment. While heralded as part of the comparable worth movement, it is not unlike union pressure to restructure wage rates at General Motors.

Interestingly, this pressure to establish pay equity is concentrated in public employment and not in private employment. Perhaps it is because there is a greater range of occupations in public employment, both male- and female- dominated occupations appearing there more frequently than in private employment. Discontent over the wage structure for male- and female-dominated occupations is more likely to arise in an industry with both types of occupations than it is in an industry dominated by either male-or female-occupations. Public employment also provides opportunity to bring public pressure as well as employee pressure to bear upon public employers.

Presumably the pay equity efforts of public employers will lead to revalidated job evaluations and a wage structure that gains acceptance by the affected workforce. The impact on the societal gender earnings gap is hard to estimate, however, since public employment is small relative to the total workforce and since the definition of pay equity likely will vary from one public employer to another.

What about proposals to establish a comparable worth definition of wage discrimination within Title VII? This is more difficult to analyze lacking any clear definition of worth. I have argued that there is no single measure of job worth and that job evaluation rests ultimately upon the values of the affected workforce. Two jobs or occupations may be valued quite differently in different work organizations. A charge

of lack of validity in job evaluation must be limited to a single employer and no system of job evaluation can be generalized among employers. Thus, it is not clear that any overall restructuring of occupational wages in the U.S. is likely to emerge from revalidation of wage structures employer by employer. And, thus, effects of revalidating job evaluations of different employees upon the gender earnings gap are impossible to predict.

The widely publicized pay equity settlements among public employers across the country from New York state to Minnesota and to Los Angeles certainly are related to the comparable worth phenomenon but must be distinguished from it also.

These settlements, either legislated or negotiated, focus upon potential wage restructuring for occupations in a single employer's workforce. That workforce may be city-wide or state-wide, but only the wage structure for a single employer typically is affected.

Pay equity determination in New York state is not unlike revision of job evaluation and wage structures at AT&T. Whenever a wage structure fails to meet the norms and expectations of a sizeable proportion of an employer's workforce, that job evaluation is revised and revalidated.

Pay equity is slightly different in so far as the attention and concern for comparable worth has advanced occupational wage comparisons to a higher level of visibility than twenty years ago. Also, since most of the pay equity settlements have occured in the public sector, that visibility and attention has aided in achieving political pressure for a change, a pressure absent in the private sector.

In a very real sense, the comparable worth movement has gone mainstream by directing change for pay equity at the level of the employing institution rather than through sweeping national legislation. The switch from comparable worth to pay equity, and the switch to mainstream tactics for wage restructuring probably are accomplishing more change than otherwise would have been the case. The focus has changed somewhat, however. The underlying motivation and rationale now rests more clearly upon traditional pay equity concerns, an occupational group of employees charging unfair treatment and arguing for advancement in the wage structure. Broader social concerns for poverty and the relative earnings of all women in the workforce have less meaning.

The focus also has shifted somewhat from alleged discrimination under Title VII to direct employer pressure to accomplish wage structuring. As a "civil right," comparable worth may be a looney idea, but as an employee interest to be sought in collective bargaining or through other pressure, "pay equity" is no longer loonier or less legitimate than any other wage demand. One side consequence of the whole effort has been the re-examiniation of job evaluation and the recognition that validity of the process is determined through acceptability of the outcome. As a process for relative wage determination, job evaluation is still as justifiable and effective as it was twenty years ago. Social

norms and values, not job evaluation, are the key determinants of occupational wage structures of employers. And as social norms and values change, so will the occupational wage structure resulting from job evaluation.

Where does this leave us?

Note, first, a distinction between the issue of a greater earnings gap (which is cited as the motivating cause) and the issue of occupational wage differentials which is the immediate focus of the comparable worth controversy. The gender earnings gap is a social issue. Wage discrimination, the focus of comparable worth, is an employer issue. An identifiable defendent must be cited in any Title VII charges. The earnings gap focuses upon worth of persons, women are paid less than men and, perhaps, equally qualified women are paid less than men in society. The comparable worth controversy focuses upon the relative worth of occupations, not persons, in a specific employment setting. We are comparing the worth of persons in society with the worth of jobs or occupations in an employment setting. Comparable worth and the gender earnings gap are doubtless related, but the logic of the two issues is quite different and the relationship becomes quite strained. Either comparable worth payment of occupations or elimination of the earnings gap might be achieved through different means without noticeable effect on the other. Indeed, many would argue that we have already achieved comparable worth payment of occupations.

Although the comparable worth doctrine is gender neutral, the comparable worth phenomenon is a gender issue, and is directed primarily at the economic power of women in society. It is basically a social and political movement aimed at redistribution of power among the genders and approaches it through redistribution of occupational earnings. Relative occupational earnings change only as value orientations change and the movement ultimately must change values of the workforce to succeed in restructuring relative occupational earnings.

Some fear that the comparable worth movement will result in a nationally imposed system of job evaluation, one that prescribes either a single technique for job evaluation or that prescribes relative occupational wage differentials. I find this prospect unlikely. Wage regulation is accepted during war or other extreme times, but is unlikely otherwise. In fact, job evaluation was required by the War Labor Board in the 1940s to justify changes in wage differentials. No single approach was prescribed; all that was required was some form of evaluation as rationalization for wage changes. A national system of job evaluation was imposed in the Netherlands following WWII as a means of controlling inflation but did not last long--necessary supplies of labor to certain occupations were not forthcoming at the prescribed rates and the system soon fell apart. Acceptance of job evaluation results is the key test of validity of job evaluation and the willingness of workers to accept employment at the prescribed rates is a test of validity.

Concerns about potential bias in job evaluation have been raised and there doubtless will be greater efforts to prevent usual sources of bias. I might note that the research to date does not indicate any general tendency for job descriptions or evaluations to display bias related to gender of the analyst or gender of the job incumbent. Research does indicate a general tendency to attribute less worth to so called women's occupations, but those are the occupations already lower paid. Attributions of occupational worth appear to reflect either traditional earnings differentials or gender domination, both of which are correlated. The major source of bias appears to reside in general social values rather than in the process of job evaluation. Efforts to establish so called pay equity through legislated job evaluation or through employee rejections of established job evaluations may well alter pay relationships within employer establishments, but it is unlikely that any major alteration of industry pay differentials will result from this. And so the effect on a societal gender earnings gap is incalculable.

Too many, in my opinion, see the comparable worth movement as an attack upon techniques of job evaluation. The ideal system of job evaluation at best merely captures and mirrors the social judgements of worth held by the effected workforce. Only as those social judgements change will the results of job evaluation change, which is why I perceive the comparable worth phenomenon to be primarily a social/political effort. It seeks change in values which tend to attribute less worth to so called women's occupations.

Ultimately, this change in values, if accomplished, probably will impact on wages more through changed labor supply behavior than through job evaluation. The job evaluation impact will be limited to single employers, while refusal by women to work in that they consider to be low wage occupations will more likely bring about redistribution of employment and reduction of the earnings gap. In fact, there is some evidence that this is occurring already--the gender earnings gap is less for younger women than for older women. We don't know whether or not this reflects change in employment patterns, wage rates, or perhaps merely similarity of cohorts at young ages which disappear with maturity. Let's be optimistic and assume that it reflects real change which will continue. Looking off into the future, I can foresee a time when the comparable worth movement has faded--yet the comparable worth doctrine will still be vital. There will be concerns about comparable worth, but they will not be associated with gender.

The basic issue of comparable worth may not have changed much since then but the logic of the response has changed. Job evaluation, in a real sense, now provides the logic of response. And that logic is tested continually in every employment setting. Keep in mind, however, that job evaluation and comparable worth are employment issues and that a gender earnings gap is a social issue. The comparable worth movement has blurred this distinction. The comparable worth doctrine is applied today within employment settings, yet the gender earnings gap remains. The search for equity in both the employment context and the social context will continue through history. True equity likely will never

be achieved. Social norms (and thus equity definitions) change over time. And justice and equity in one context need not relate to justice and equity in another context.

Relative pay equity for jobs in an employment setting is achievable and we know how to achieve it. I'm not sure that we know yet what earnings equity for people is or that we could achieve it in ways other than free access to jobs. I do, however, think it is important to keep the distinction between pay equity and earnings equity clear.

So much for this analysis of comparable worth. What you have heard is an analysis of the comparable worth movement by an ardent advocate of the classic comparable worth doctrine. One can endorse the doctrine without endorsing the objectives of the movement. At minimum, the movement has served a useful purpose in forcing re-examination of social norms and values, whether traditional norms and values are revalidated or changed. For that I am thankful.

* * *

REPORT ON THE IPMAAC JOB ANALYSIS PROJECT

Chair: Ronald A. Ash, University of Kansas

Participant: Bruce W. Davey, State of Connecticut

The IPMAAC Job Analysis Task Force, appointed by the IPMAAC Board of Directors, May 1983, is developing a set of competency based standards for personnel assessment specialists in the field of selection, performance appraisal, training needs and program evaluation, job analysis and organizational effectiveness. The results have been analyzed by Jack Lawton, University of Michigan and Ronald A. Ash, University of Kansas.

Ron Ash presented a summary of the cluster analysis of the data contained in task analysis questionnaires completed by 447 persons. The result is a comprehensive set of ratings for 217 tasks and a cluster analysis grouping these tasks into 15 clusters. Each empirically derived cluster is composed of a set of relatively homogeneous tasks. The summary of the 15 task clusters are presented in the following table:

Summary of Cluster Analysis Results of the IPMAAC Personnel Assessment Specialist Task Inventory Data

	Personnel Assessment Specialist Task Clusters	Cluster Ratings: Proportion of Job	Cluster Ratings: Mean General Importance
CL1.	Job Analysis, Description, and Classification Activities (17 tasks)	15.5%	28.8
CL2.	Selection Procedure Development Activities (21 tasks)	13.2%	30.6
CL3.	General Personnel Assessment Management and Supervisory Activities (23 tasks)	12.9%	31.9
CL4.	Training and Education Activities (34 tasks)	11.6%	20.2
CL5.	Information Exchange and Communication Activities (11 tasks)	10.9%	22.6
CL6.	Technical Personnel Assessment Management and Supervisory Activities (13 tasks)	6.5%	32.5
CL7.	Equal Employment Opportunity, Affirmative Action, and Related Activities (17 tasks)	6.3%	23.2
CL8.	Selection Procedure Validation Research Activities (22 tasks)	5.9%	24.7
CL9.	Basic Test/Assessment Procedure Administration Activities (10 tasks)	3.7%	23.6
CL10.	General Data Analysis Activities (8 tasks)	3.1%	24.8
CL11.	Applicant Evaluation and Screening Activities (7 tasks)	2.7%	27.6
CL12.	Recruitment and Preliminary Screening Activities (10 tasks)	2.4%	18.7
CL13.	Assessment Center Development and Management and Supervisory Activities (6 tasks)	2.1%	27.7
CL14.	Non-Personnel Assessment Management and Supervisory Activities (8 tasks)	1.9%	24.0
CL15.	General Personnel Research Activities (4 tasks)	1.7%	26.5

* * *

MICROCOMPUTING IN PERSONNEL (Symposium)

Chair: Larry S. Jacobson, State of Connecticut

Introducing Computer Applications Into Organizations

Donald Harris, Metro-North Commuter Railroad,
State of New York

The introduction of computer applications into organizations is an important topic because applications in the human resource area are proliferating. These applications have become relatively cheap, manageable and familar. Organizations will increasingly turn to these applications. Many computer applications introduced into organizations are failures. Introducing computer technology introduces change into an organization: changes in procedures, jobs, patterns of interaction, values, people, etc... This change must be managed properly if both the application and the organization are to be successful. There is no general agreement as to who is responsible for managing the process, or as to how it should be done.

Typical problems encountered in introducing computer applications into organizations:

- lack of support of key manager or resource allocator needed to get going
- can't find the time to determine the requirements adequately
- the technical people can't understand the non-technical people, and vice versa
- users who are uninvolved or uninterested in the application
- consultants and vendors who inflate their promises and capabilities
- applications which prove to be too limited and inflexible after being implemented
- the goals of the application, as perceived at different levels of the organization, are incongruent
- ownership of, and responsibility for the introduction of, the application are poorly defined
- impact upon organization not planned for, leading to undesirable consequences

Some recommended strategies and steps in introducing computer applications into organizations:

1. Specify and gain agreement as to the nature of the problem(s) with the status quo, which has prompted an interest in a computer application

   - describe the problem very specifically and clearly, including why it is a problem

   - see if others agree, particularly at other levels and areas of the organization

   - be prepared to redefine the problem if broader support is desirable

   - secure agreement both as to the nature of the problem and the criticality of the problem. Is solving the problem a priority? for everyone?

   - package or present the problem in an appealing, perhaps humorous, but above all, understandable manner

2. Develop a general understanding of how a computer application might address this problem

   - find out what's available, what others in the field are doing, what products or services vendors are selling, what the DP/MIS department might be able to do

   - determine feasibility in the broadest sense, ballpark a solution to the problem

3. Determine who in the organization would be impacted by the application, and involve them in all subsequent planning and implementation

   - from indirect users you want involvement; from direct users, commitment

   - "users participation" may range from occasionally consulting with the users, to having them participate in key decisions, to having them pretty much in charge of the whole process, and also needs to be defined if it's promised.

   - methods of getting users to participate include surveys of attitudes and expectations, demos, interviews, and "project teams".

   - the ownership issue is critical: ownership builds commitment and the exercising of responsibility

- who should be responsible for introducing the application?

4. As part of planning, assess the organization and what impact the application will have on it

   - where is the organization on the computer maturity or experience continuum? This may set certain constraints on the application

   - where is the organization on the democratic/authoritarian continuum? This may constrain introduction of the application

   - are there any significant impacts of the application on who holds power in the organization? Will it bring changes in status and influence?

   - are there any significant impacts of the application on personal relationships and patterns of social interaction in the organization?

   - who will bear the costs of introducing the application compared to who will reap the benefits of the application? A discrepancy here indicates trouble

5. Determine the needs or requirements that the application is being introduced to meet

   - if you don't determine your needs or requirements, you have no standard by which to judge which of several alternative applications is best, or, if you have already chosen one of these alternatives, whether the application is successful once its implemented

   - needs or requirements should be articulated very specifically and comprehensively

   - to develop a statement of your needs you might consider, at least in a general manner, the inputs and outputs of your application, the volume of data that might be involved, the number of transactions required, time requirements, the abilities and training needs of the users, documentation requirements, the need for edits and audit trails, security issues, back-up requirements, the type of maintenance or application support available, the degree of flexibility needed, etc.

6. Identify possible solutions/applications to meet your needs

   - contact and visit others using these applications, discuss their satisfactions and dissatisfactions, ask them for their assessment of whether the application can meet your needs

- where possible, arrange for a demonstration of an application

- a matrix of possible solutions/applications against a listing of your needs and requirements may be helpful

- if a package from a vendor is involved, consider the vendor's reputation, profitability and growth, the service and training provided, along with the product itself and its cost

7. Select the application that best meets your needs

- the "best" application, from the point of view of meeting your needs, may cost too much, require hardware your organization is unwilling or able to buy, be incompatible with other applications the organization has or is planning to acquire, and/or be unproven.

- select the application that best meets your needs under the circumstances

- try to resolve differences of opinion as to which application or alternative should be persued. Once again, ownership and commitment are critical

8. Implement the application

- depending upon the scope and size of the application, implementation may require more planning, effort and time than all of the preceding steps combined

- if it's a large application or system, you may want to go through some type of standard MIS life cycle methodology, including steps such as:

    - the development of a functional specification, or technically-detailed statement of what the application is supposed to do

    - system design, or the largely technical development or adaptation of the application

    - data collection, data conversion and data input

    - the development of manual procedures to be used in operating the system

    - implementation on a test basis, while maintaining in parallel the prior way of doing business

    - user training

9. Assess the introduction and implementation of the application

   - have some formal procedures to do this, a requirement that an evaluation be performed
   - how do you know if the application is a success?
   - what did you learn from the process, that you could use the next time around? How would you do it over?

10. Some general recommendations:

   - don't be passive in the process, or assume that someone else is addressing the difficult issues. You may end up as more of a victim than a user
   - ask that data processing terminology be explained to you in terms that you can understand
   - at a minimum, learn enough about computer technology to avoid being totally dependent upon technical people
   - the technology is moving steadily in the direction of facilitating your independent use and control of it: end-user computing, end-user application generators, natural languages, voice-activated systems, artificial intelligence, etc.
   - get and stay in touch with what's going on in the field of human resource information systems (HRIS)
   - explore developments in the field of organizational change
   - explore developments in the field of systems analysis and design. It is an exceptionally diverse and wide-open field.
   - finally, in your spare time, keep up with the constant, rapid and in many ways revolutionary changes going on in computer technology and computer markets

* * *

Problems and Payoffs in Automated Applicant Tracking

Glenn G. McClung, Denver Career Service Authority
Denver, Colorado

About two weeks ago I was cleaning my desk, and finally threw away an item I bought second-hand for my first job in 1953 as an Electronics Apprentice. It was a Pickett pocket-size magnesium slide rule. I don't know why I kept it all these years, except that it had my scratched-in formulas for things I couldn't remember then and still can't.

In 1960, when I took my first professional job with the California State Personnel Board, I was still using that slide rule and did for a number of years more, adding more scratched-in formulas as I went along. At any rate, sometime in the mid 60's the pocket calculator came on the scene, as did the IBM 1401 Computer for our shop in Sacramento. Now the IBM 1401 was a fancy machine which, although rather large, had almost as much internal computing power as the Timex/Sinclair my wife bought me for Christmas in '82! But frankly, the change in the office did not seem that revolutionary. We'd had a data processing section with punch-card equipment for years, and the computer was basically just a way of doing things faster and better. While it was true we'd previously done things like item-analysis and statistical analysis by hand, automated test scoring and semi-automated candidate notices were not new ideas. Having grown up in one of the largest personnel departments in the country, it wasn't until I went to Denver in 1972 that I fully realized the gap in technology between large and medium to small personnel jurisdictions.

As I recall, in '72 we still had a test scoring machine in Denver that involved reading the score from a needle on a dial, kind of like a thermometer. Within a short time, however, we moved on to one of the newer optical-scanners, skipping the mark-sensing machines entirely. Our biggest problem technically was our lack of capacity for any sophisticated and timely test analysis. What analysis we did was either by hand or, later, on a Canon desk-top printing calculator. While a "black-boxed" attachment was presumably available for our scanner, the cost was high and was virtually impossible to justify simply on the basis of more sophisticated test analysis. After all, we could ship things to the centralized data processing shop for our city. And assuming we could explain what we needed to our folks at Data Services, we might eventually get something back.

Frankly, an even bigger problem for us, and one which was far easier to explain, was the cost of slowness of the routine clerical operations involved in the total recruitment, scheduling, testing, scoring, candidate notification, and certification process. Those of you with multiple-phase examinations know just how much time can be involved. In our case our volume and backlog was such that we couldn't produce final test results until a week or two after the last test phase was complete. This problem eventually got so bad that we put three clerks on a permanent 4 to midnight shift just to cope with the problem of assembling test data and getting out results notices. This was 1980.

By 1980, our agency had had a lot of experience dealing with our centralized data processing shop. In fact, through remote terminals, we had had an automated position control and complement control system for about eight years. Unfortunately, even after eight years, that system had never become reliable enough to abandon the manual system it was supposed to replace! It was, in fact, a giant albatross, which no one knew quite what to do with. By 1980, after about three years in the making, we also had a main-frame based ethnic census of exam competitors that gave us gross figures once a year for our Affirmative Action Plan.

All in all, I was not impressed with the service available from our City Data Services Division. The exam program needed automation, but not, in my opinion, another attempt at a main-frame application.

In 1980, we became aware of a commercial mini-computer application called "TRAC". It was a comprehensive applicant tracking system which had been specifically designed for the intricacies of government merit systems, and was actually in use in several places. While a number of firms were marketing what I've since heard Bruce Davey call "vaporware", TRAC looked real. One of the early installations was San Bernardino County, so George Nelson of our staff went out to visit with Ted Darany. We were sold, and by mid-1981, had a specific proposal moving through the budget process.

While TRAC has been implemented on main-frame systems, one of the most attractive features to us was that it was basically designed for stand-alone, user-run, mini-computer applications. Most of us in the examining business have good reasons for not wanting to give up control of our exam data or the machinery necessary to run the selection program. Budget people, however, are a little less understanding. Also less than understanding are the folks in central data-processing, who often have millions invested in their main frame equipment and need business to stay viable. Needless to say, our first step toward implementation was salesmanship.

We pitched our system, not for program improvement or advanced research capabilities, but simply on cost benefit. The scheduling and processing of candidates in a merit system can be an expensive, labor-intensive operation. In 1981 I had about 23 clerical positions in my Division. I offered to trade four of them for TRAC, and was able to show that TRAC would more than pay for itself within five-year lease/purchase period. Our package, by the way, included a Prime Mini-Computer with 64 megabytes of hard-disk drive, Diablo letter-quality printer, three CRTS, and a new Scantron Test Scoring Machine. Hardware costs were about $80,000 and TRAC software, including installation and training was about $30,000, including installation and training. Insurance, maintenance, and carrying charges brought that to about $150,000 over a five-year period. While the City later decided to buy the system outright, the lease/purchase argument made it easy to compare cost manpower. Things looked good, and we expected to have the system on board by January of 1982. Unfortunately, there were more problems to come.

Facing the need to replace the expensive and disastrous complement control system I mentioned earlier, my agency and the budget division decided to hire outside consultants to review the EDP needs of the total personnel system and prepare an RFP on the total package. Throughout the study, my division emphasized the essential independence of the examining program from other personnel functions, and succeeded in securing a final RFP which made it possible for a firm to bid all or part of the package. As we had expected, no one but TRAC presented bids that could meet specifications for our applicant tracking needs.

The TRAC stand-alone system won the contract for our employment program, with the rest of the personnel system going to a large software consulting firm, for a main-frame application. As a matter of side-interest, the City eventually had to break the contract for the Personnel System implementation, which turned out to be "vaporware", and just last month started to install yet a third system! TRAC installed in mid-82 and has been functioning very well since.

The many months which preceded the arrival of our in-house minicomputer were well spent. One of the major problems in a computer installation is the system it replaces. People are used to doing things a certain way, and there are often little sub-procedures that don't even make sense. We had reviewed our procedures and revised most of our forms before TRAC arrived, and had prepared our clerical staff for the coming change. While system modifications in software are always necessary to fit the particular installation, we had resolved to change our system, where we could, rather than laying too many demand on the vendor.

After TRAC's arrival, about a week was spent on training, as offered by the vendor. But we still weren't ready to go. At least a week was spent just having staff get familar with the program and equipment, including some of the games with which the computer came equipped. Another couple weeks were spent simulating and modeling actual exams, rather than running the real thing. We were committed that we would not run a "dual" system once we got underway. When we did start, we didn't try to convert all exams at once, but phased them onto the computer over a several month period. Some of our more complex clerical exams were purposefully not converted for more than a year. Another strategy was to appoint a competent technical person as "system administrator", even though routine operations are all performed by the clerical staff. Where stand-alone systems have not worked well, it is frequently because of turnover or a void in this type of technical support.

Our system has been extremely successful and has been expanded. Since our workload has gone up while our staff has contracted, I doubt we'd have made it the last three years without TRAC. In addition to applicant tracking, we are also using our in-house mini for salary survey processing and fringe benefit studies, which saves us time and money in relation to using the central mainframe. We've added three terminals, a second printer, and 80 megabytes of additional disc s storage, which gives us room to move into the word processing of our written tests.

As you can tell, I really don't need my slide-rule anymore.

* * *

Microcomputers For Conferences and Networking

Patrick T. Maher
Personnel and Organization Development Consultants, Inc.
LaPalma, California

Even the simplest of computers possesses the ability to communicate with virtually any other computer, to access computer capabilities far in excess of its individual capabilities, to obtain the power of mainframes, and to exchange information. Too often, individuals purchasing their own microcomputers fail to fully realize the potential in even their "home" computer, and believe that other potential uses are reserved for more expensive "business" computers.

System Requirements

Any system starts with the basic hardware known as the "computer." In addition, a device commonly referred to as a "modem" is required. The modem connects the computer and the telephone, thereby allowing the phone lines to transmit information from one computer to the other. Modems, like computers, come in a variety of types, capabilities and styles. "Smartmodems" are capable of dialing numbers and automatically connecting onto the system and automatically answering incoming telephone calls from other computers. These modems are generally expensive, costing in excess of $500.00. Cheaper, albeit simpler, modems are available, starting at under $100.00. Such modems require that the user perform all functions and the modem serves as little more than an extension of the basic telephone. In addition, the computer must have a serial (RS-232) port. If such a port is not available, then either one must be installed or the computer must be modified in some manner. With the growth of communication among computers, many microcomputers are including a modem as a built-in component. In addition, at least one manufacturer sells a telephone that contains a modem. There is little doubt that in the near future modems will be a standard piece of equipment on any microcomputer.

Special Interest Groups

Special interest groups (SIGs) are exactly what the name implies: a group of persons who have some special interest in common. Historically, SIGs were computer buffs who joined together to exchange ideas and experiments in hardware and in software.

More recently, SIGs have expanded to include such diverse groups as real estate investors, attorneys, farmers, and others.

Bulletin Boards

Bulletin boards (BBs) are specialized information systems maintained for the purpose of exchanging specialized information. Many are operated free of cost, and the only thing needed for access is the telephone number of the BB. Others are limited to members of a particular SIG for whom the BB is operated.

BBs usually have only a single access line, requiring the caller to pay the standard long distance toll charges. Access is also limited, allowing callers to leave or read messages, but not allowing for real-time exchange of information.

Anyone with the basic equipment and proper software can start a BB. Because of the ease with which they can be operated, it is estimated that over 4000 BBs are operational in the U.S.

Electronic Mail

Electronic mail (EM) refers to the use of computers to send letters, memos, or any other type of document from computer to computer rather than through the postal service. The advantage of EM is that messages can be transmitted and received immediately and are not dependent on regular mail delivery. A hard copy can be printed if there is a need for a permanent record, although legal documents requiring signatures or other legal notations must have a hard copy sent separately. EM can be transmitted either through a computer service that offers the service, or directly between systems that have the capability of automatically receiving and storing the information until it is called for by the receiver. EM is probably more expensive than standard mail delivery, but probably cheaper than the cost of express mail services. This is especially true for systems that can store messages and send them automatically late at night when rates are cheaper, if the recipient's system automatically receives and stores messages also. Personnel professionals, as well as any other individual, can use EM as a means of transmitting and receiving a variety of information and messages.

Networking/Conferencing

Microcomputing can be used for both networking and conferencing, either by direct communication between computers or through an online service. Networking can also be accomplished through various SIGs and BBs, as well as directly between users. When an on-line service is used, conferencing can involve virtually an unlimited number of persons. Participants type in responses and read information from others in real-time. In addition, it is possible to share information in existing files if desired.

On-line Services

Currently, there are two major companies offering on-line services with access to mainframe computers and a variety of services: Compuserve and The Source. Both are relatively the same in terms of services, but Compuserve tends to be a little less expensive. In particular, The Source has a minimum monthly charge, while Compuserve has a minimum charge only if prime-time service is desired. Both services provide electronic mail service, a variety of BBs and SIGs, and the capability for real-time exchange of information for conferencing.

The on-line services charge an hourly fee, but access is through a local telephone number. If the on-line service is used for conferencing, the hourly fee is no more than for the same amount of time spent on long distance telephone connections, but is more efficient since numerous participants can interact. Other specialized services exist, such as WestLaw computerized law library, and a variety of computerized library services. These services require subscriptions for access and can be quite expensive, although also quite valuable for the specialized user.

Exchange of Information

The various systems described above can be used to exchange information among individuals. For example, if one person desires a test from another person, the text file containing the test can be transmitted via computer, providing instantaneous access to information. Data bases of one organization can be made available to another organization. Programs operated by one system can be made available to other users by either exchanging the program via modem or by using one computer as a terminal to access the computer with the program. At the present, many personnel professionals are members of various consortiums, participating in item banks and otherwise sharing information, ideas, and resources. With the increasing use of the microcomputer, personnel professionals will find one additional resource to share.

Working At Home

There is an increasing trend for many types of jobs to be done at home through the use of the microcomputer. Merely having a modem capability changes any microcomputer into an extension of the office. Many tasks performed by personnel professionals, such as item writing, preparing job analyses, and numerous other types of reports and studies could just as easily be done at home and transferred to the office system for final printing.

Summary

The uses of microcomputers are limited only by the imagination and knowledge of the user. Simple and inexpensive hardware can greatly expand the capabilities of even the most basic computer system.

* * *

CHANGE IMPLEMENTATION TECHNIQUES FOR PUBLIC INSTITUTIONS (Symposium)

Chair: Michael L. Pendergrass, Montclair State College

The Use Of "Stakeholders" In The Development Of A Selection Procedure

Katherine W. Ellison, Department of Psychology, Montclair State College
Upper Montclair, New Jersey

Police departments traditionally have been the focus of attempts by a variety of interest groups to assert political control. The fight for "civilian review" has been endless, and lawsuits alleging discrimination in hiring have been common. For these, among other reasons, police often have been suspicious of the gentle ministrations of outsiders.

Also, police selection procedures have often been haphazard, bases either on political whim, or on a parody of valid procedures: that is, on "psychological tests," such as the MMPI, which have not been specially validated for policing. The assumption is commonly made that the stereotypical "normal" person is the best candidate for a police officer, an assumption contradicted by the existing evidence (Lefkowitz, 1977). It is extremely rare to find a job analysis for a police department which establishes valid criteria. Indeed, in a job such as policing for which (in common with many "service" jobs) performance criteria are often ephemeral, the question of appropriate dimensions for a job analysis present special problems.

In the spring of 1983, I was asked to develop a new selection procedure for a medium sized police department, with an authorized strength of about 100 officers, in a suburban town with a population of 40,000. This town has a varied ethnic and socioeconomic mix, and a history of sensitivity to racial issues. Litigation against police selection procedures has been common.

The selection procedure which had been used previously was a hodgepodge of "personality" tests; it had been developed with some attempt at concurrent validity, but without a comprehensive job analysis. That it lacked face validity was obvious from the disparaging comments of officers who had been involved in the validation process.

The job analysis which was to form the basis of the new selection procedure began with some standard measures of performance: number and type of calls, time spent in various activities, and the like. Hundreds of hours were spent analyzing records and "riding along" to observe performance first hand. These, however, do not tap the more critical QUALITATIVE dimensions of police performance. For this, the decision was made to include not only those directly invloved in evaluating the performance of patrol officers, such as supervisors and the incumbents of the job, but also a variety of groups who might be considered to have what Edwards (1980) has called a "stake" in the job. This way, the demands of citizens for involvement could be met in a way that was consistent with sound assessment practice.

Groups identified as "stakeholders" included officers at all levels of the department, elected and appointed officials of the town, opinion leaders, such as clergy and press, people who have special contact with police, such as the ambulance squad, adolescents (the group most likely to come into contact with police), and a random sample of citizens. (We even interviewed arrestees, and the proprietor of the luncheonette where officers congregated.) The instrument used to assess what these people felt was important for police in their community was a modification of that developed by Dunnette and Motowidlo (1977). Dimensions included:

a. Crime prevention

b. Using force appropriately

c. Traffic maintenance and control

d. Maintaining public safety and giving first aid

e. Investigating, detecting, and following up on criminal activities

f. Report writing

g. Integrity and professional ethics

h. Dealing constructively with the public

i. Handling domestic disputes

j. Commitment, dedication, and conscientiousness

k. Teamwork

Interestingly, officers rated the dimensions of crime prevention, public safety, dealing constructively with the public, and integrity higher than did citizen groups. (Not surprisingly, they also rated "teamwork" much higher than did any other group.) Citizens and politicians were higher than officers on the importance of the dimension "using force appropriately."

Dunnette and Motowidlo's formula was used to assess these dimensions: they were combined with those identified by the more traditional job analysis strategies. The final selection procedure involved a series of stages.

a. Written examination, tapping reading comprehension, basic arithmetic, judgement, memory, understanding diagrams;

b. A physical fitness examination;

c. An "oral board"; a structured interview;

d. Appearance;

e. Emotional stability;

f. Overall suitability for police work.

The composition of the oral board also reflected the commitment to the involvement of "stakeholders:"

a. A sergeant or lieutenant from the Department, chosen by the officers;

b. A patrol officer from the Department, chosen by the officers;

c. A command level officer from another department;

d. A psychologist;

e. A township citizen.

Analysis of this process has revealed that, while minority candidates did less well on the written test (a finding common with other research), they performed better than white males on the "oral board."

The correlation between performance on the written test and police academy standing was r=.89. Also, candidates chosen by this process have had higher ratings by both supervisors and peers at the end of their probationary period than officers chosen by other procedures, with a correlation between "oral board" rating and supervisor rating on "overall suitability for police work" of r=.64. "Rejected" candidates who subsequently obtained jobs with other police agencies have almost uniformly received low ratings in their first year with those agencies, and have been significantly more likely than the average to leave or be terminated. (All of these candidates passed standard "psychological" screening procedures.)

The use of "stakeholders" both from inside and outside the Department to assess the important dimensions of policing had a number of peripheral benefits. In addition to providing data for the job analysis, it gave the town and Department administration a better feeling for the kinds of services citizens wanted from police. It also gave the "stakeholders" a feeling of participation, which may have played a large part in the general acceptance, not only of the selection procedure itself, but also of the officers selected by it.

References

Dunnette, M.D., & Motowidlo, S.J. (1976) Police selection and career assessment. Washington, D.C.: U.S. Government Printing Office.

Lefkowitz, J. (1977) Industrial-organizational psychology and the police. American Psychologist, 32:346-364.

The Role Of Implementation In Personnel Management- Connecting Theory To Practice

Herbert Sherman, The William Paterson College of New Jersey, Wayne, New Jersey

Several studies of personnel managers have demonstrated that given the extent of information available on the viability of training techniques, personnel managers' opinions as to the utility of these techniques is highly correlated with training and development theory (specifically learning theory). However, numerous studies of the actual frequency of the use of instructional methods demonstrates that personnel managers implement programs which, in many cases, are counter to their own concepts of proper developmental theory.

This gap between theory and practice, the difference between personnel managers' attitudes and behaviors, has historically been rationalized as personnel managers' inability to implement learning theory due to the "cryptic" nature of research in the area of human development. Several writers believe that these managers were incapable of transferring this material to the business training and development environment and therefore assumed that personnel managers practiced without the benefit of learning theory. Other writers suspected that Personnel Directors were highly influenced by "fads" and therefore many perprograms were developed based upon the directors' perceptions of the social and industry acceptability of the techniques employed. Both explanations provide partial insights into the actual problem faced by most personnel managers, implementing personnel theory, yet fail to address the problem in a systematic manner by analyzing "implementation" as the key to the success or failure of any personnel program.

Implementation is defined (by Sabatier and Mazmanian, 1980) as the carrying out of a basic policy decision, usually formalized into a program or project. A successful implementation is determined by how well the division or department achieves program or project objectives and how much of the program or project has been incorporated in the organization's operations. A successful implementation, therefore, changes within the division or department to meet the objectives of the policy decision.

Programs, policies, or products however do not structure implementation processes and therefore our attention must be focused upon the stakeholders or actors who construct and implement the organizations' objectives. We must also examine the background factors or "givens" (factors which describe organizational and environmental characteristics) which define and affect the actors' levels of interactions in that these factors determine both the constraints to and forces for successful program implementation. These background factors of the implementing department

or division can be considered an aggregate "cause map" of all of the behaviors of the stakeholders and include:

A. Organizational

1. general socio-economic conditions to target population (community, wealth, education, stability)

2. organization's history of successful implementations

3. presence of professional activity (is there an affiliate professional organization?)

B. Department/Division

1. diversity (# of difficult tasks, specialists, technologies)

2. formalization (rigidity and specificity of roles)

3. centralization (narrowness of chain of command)

4. staff input into decision-making

5. excess resources

6. professionalization of staff

7. employee unionization.

Many organizations do not possess the capabilities to solve or meet the issues or needs defined by personnel managers. The solvability of the problem, that is the company's ability to utilize viable theories to changing conditions, affects the implementation process in that inadequate theories will not produce appropriate changes in the target populations nor may it alter department/division operations. Factors affecting the solvability of the defined problem include:

1. availability of valid technical theory and technology

2. diversity of target group behavior

3. target group as a percentage of the entire company

4. extent of behavioral change required.

Assuming the company possesses the capabilities of solving a specific personnel problem, one must examine the company policies developed in order to determine if the policy to be implemented incorporate the appropriate factors necessary to achieve program objectives. In numerous instances a department may have the ability but not the

facilities to implement a program properly. Personnel managers implementing programs should therefore examine the following factors:

1. validity of the causal theory in the program
2. unambiguous policy directives
3. financial resources for program
4. hierarchical integration (how many program approval points?)
5. decision-rules of implementing department/division (does the policy impose guidelines or rules?)
6. recruitment of implementing personnel (are implementing personnel committed to the program? Is department/division oriented against program?)
7. formal access by outsiders (can target population modify program?)

The last set of factors, variables directly affecting implementation, attempts to gauge the political environment and support that the program and the department will receive and how these inputs might change over time. Numerous programs have failed simply due to the lack of interest on the part of the target group and the managers are therefore cautioned to examine the following factors:

A. Organizational

1. prior need (is there demand for the project?)
2. executive support
3. target group participation
4. general industry support
5. attitude/resources of constituency groups
6. publicized attention to problem

B. Department/Division

1. commitment/leadership skills of implementing managers
2. training of implementors
3. support of department/division head

4. delays

5. monitoring of implementation

6. power to implement program.

The implementation process therefore is a highly complex process and should not be taken lightly by practitioners or academicians in the discussion of connecting theory to practice. Many "theories" fail because they are not properly incorporated in personnel policy, are contradictory to the structure and/or values of a company or department, and do not have support from either top management or implementing personnel.

* * *

A Study of MMPI Use in Police Officer Screening

Rebecca M. Baybrook, City of New Orleans Civil Service Department;
and Penny Dralle, LSU Medical School

This paper is a report on certain aspects of a comprehensive study of police recruit screening and selection in a major southern city. The comprehensive study is presented as having four basic goals 1) to determine the extent to which various components of the screening procedure contribute to prediction of job success, 2) to develop norms for the population studied, 3) to identify any influence of demographic variables, i.e. race, sex, and age, on decisions made during the screening process, and 4) to describe the various decision making processes using what the authors call a policy-capturing methodology.

The paper includes Minnesota Multiphasic Personality Inventory (MMPI) norms for this sample of police recruit candidates, analyses of the influence of demographic variables on screening decisions, and an investigation of policy-capturing for two components of the screening procedure, the police background investigation (including polygraph), and the psychological screening. Utilizations of the MMPI play an important part in all aspects of the study.

Figure one is a schematic representation of the entire screening and selection procedure. At decision points on Application, Written Examinations, Agility and Medical Examination, candidates can be rejected and no further evaluation occurs. This report is concerned with the Polygraph and Background Check - Psychological - Psychiatric part of the procedure. Background information collected by police investigators and polygraph information are combined in a police report. The police report concludes with a statement about the candidate's acceptability to the police department. That statement, however, is not sufficient for eliminating the candidate from further consideration. The police report is then sent to the psychologist. The psychologist reviews MMPI scale scores and the police report including the police

department's opinion about the candidate's acceptability. The psychologist also makes several judgements about the candidate's suitability for police work. If the candidate appears unsuitable or of questionable suitability, he or she is referred to a psychiatrist for a structured clinical interview which is used to make a final decision about the candidate's suitability. No applicants are rejected for psychological reasons without benefit of a personal interview with a psychiatrist. Both the background investigation and the psychological screening involve considering relatively large amounts of information prior to making a recommendation.

Hypotheses and Assumptions Governing the Study

Various expectations or hypotheses about police department judgements based on the background investigation and polygraph were considered. The purpose of the background investigation is to identify prior criminal behavior, particularly involving drugs. It is also an opportunity to gather information about prior job-related education or work experience, including references. We expected two factors to emerge from the information collected: one based on the polygraph and criminal records and the other based on education and work experience. We expected both factors to contribute significantly to the police department's evaluation.

The psychological and psychiatric screening aspects of the police recruit selection process function to 1) identify who have significant maladjustments and 2) develop hypotheses concerning characteristic traits or behavioral patterns in individual applicants which might significantly interfere with performance as a police officer. Characteristic traits and behavioral patterns have been related to job performance using job analysis, meetings with police personnel, and common sense.

Method

Subjects: The subjects were all 1980 and 1981 police recruit candidates who had successfully completed earlier portions of the screening procedure and were investigated by the police department and evaluated by the psychologist. The applicants were between the ages of 19 to 35, had a high school diploma or equivalent, and had passed a content valid examination, agility test, and preliminary medical screening. Three hundred and fifty six persons met this criteria. Of the personnel files for these persons, 315 contained information about their final employment status with the police department. In addition, psychiatric evaluations were completed on 124 of these applicants.

Data: Data were coded from the files of those 356 applicants sent to the psychologist between January 1980 and December 1981. Measures were chosen based on the information available in the files, hypotheses, and consideration of rater ease and reliability in coding the items. The psychological ratings were developed

based on an analysis of effective and ineffective performers on the police department. The psychiatric evaluation scales were developed based on an adaptation of Bellak et al.'s ego function assessment.

Analyses: Norms were developed by calculating the means and standard deviations of the 't' score distributions for each race and sex group. Measures used for each component of the screening procedure studied were submitted to a principle components analysis. This type of analysis generates factors which indicate variables which are highly related and therefore may be redundant contributors to decision making. It also suggests the underlying structure of the measures collected and analyzed. Stepwise linear logistic regression analysis was used to investigate the contribution of different variables to decision making. Linear logistic analysis does not require assumptions about multivariate normality. The data could not satisfy such assumptions and therefore linear regression was inappropriate. Linear logistic analysis does provide statistics which are analogous to those found in linear regression.

Results

A. Background measures. The factor loadings for factors identified by the principal components analysis of background measures explained 61.6% of the variance in these measures. All of the polygraph items loaded highly on the first factor. Factor scores on this factor can be considered a summary of the candidate's polygraph results. The second factor appears to be an arrest record factor containing both self report and archival records of suspected criminal activity. The measures related to education loaded on factor three. The remaining three factors are all related to prior work experience. Factor four contains measures of prior work experience in organizations which are similar to the police department in terms of personnel systems and political forces. Factor five is a more general factor reflecting both the applicant's age and the stability of his or her work record. Factor six contains measures relating to work experiences which were similar in responsibilities and assignments to police work.

The stepwise logistic linear regression used these six factors plus age, race and sex to predict the police departments' evaluation of the candidate's acceptability based on the polygraph and background investigation. Factors one, two three, and six made significant contributions to predicting the police evaluation. The goodness of fit statistic, analogous to a significance test of R^2, indicates that the model derived using the logistic analysis, predicts the outcome (police evaluation) significantly better than chance.

B. MMPI Norms. Applicants were compared to norms for their sex groups. Our results are similar to the results of previous studies of other police departments' applicant populations. Both profiles differ significantly from the normative profile but both are well within the clinically normal range. Moderate elevations on K, Hy, Pd, and Ma scales and a slightly lowered Si scale are typical of these police applicant groups. In addition to these expected elevations, 't'-tests for differences between the male sample and normative population were significant for L, D, Mf, Pt, and Sc. However, the mean differences were small. Our female sample did not show the typical Hy elevation and although present, the lower Si is not significant. In the female sample, slightly higher (than the norm) L and Sc means were observed. In addition, the mean Hs and D scale scores are slightly lower than the normative mean. The observed means for Hs and D scales are probably due to a tendency for the female group to respond to items related to concerns about their bodies, health, and general dissatisfaction and discomfort more like the male applicant sample than like the female normative group. Although there are significant differences between means for the normative sample and this police recruit candidate sample most of these have little practical utility in making clinical decisions.

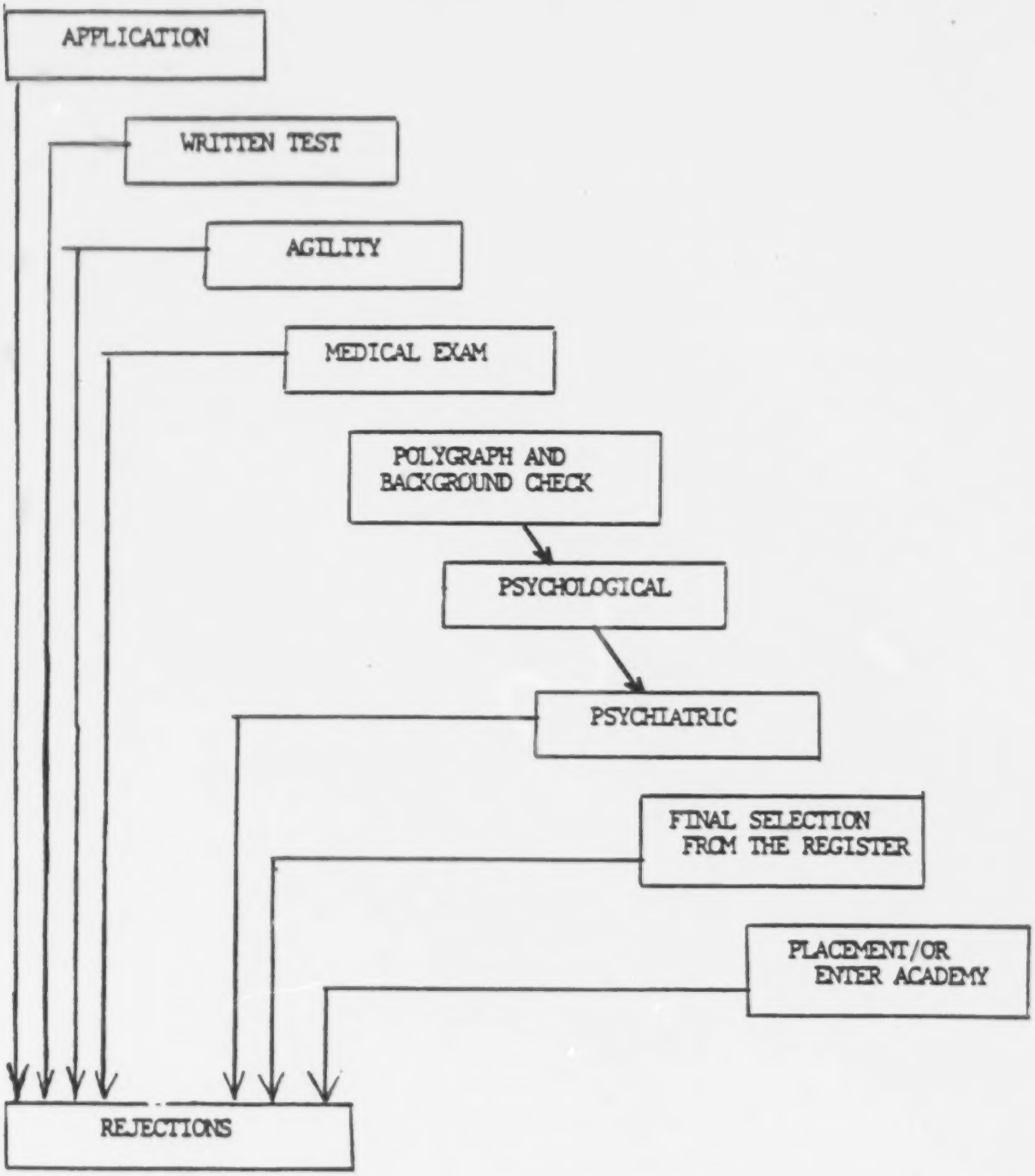
C. Race and Sex. In order to investigate the influence of sex and race, the various MMPI scale scores, were regressed on sex and race. For each of the following MMPI scales, the linear combination of sex and race significantly predicted scale scores: L, Hs, D, Hy, Pa, Pt, and Si. In addition, sex contributed significantly to prediction of scale scores on Hs, D, Hy, Pt, and Si. Race contributed significantly to prediction of scale scores on L, D, Pa, and Si.

D. MMPI Factors. A principal components analysis of the MMPI scale scores yielded five factors. These factors accounted for 70% of the variance. Factor one loads heavily on scales related to adjustments or maladjustment and is similar to Welsch's anxiety scale and Kassebaum's ego strength versus ego weakness factor. This seems to measure an obvious effort to present one's self in the best light possible without any faults or problems or at the other extreme represents willingness to acknowledge physical, cognitive, emotional, and interpersonal difficulties, concerns, or discomforts. Factor two seems to measure extroversion and is similar to Kassebaum's second factor. It appears to be a measure of social conformity in appearances and interpersonal orientation. Factor three seems to be a measure of apathy versus activity and enthusiam. Factor four contained scales indicating lower coping skills and assertiveness with high levels of tension and uncertainity.

E. Psychological and Psychiatric Ratings. The principal components analysis of the psychological and psychiatric ratings reduced the ratings to separate sets of factors even though the measurement goals were the same for the two ratings. Factor one contained the ego function ratings suggestions that such items as reality testing, judgement, regulation and control of impulses, and object relationships and defensive functioning are not differently rated based on the one hour structured interview. Factor two consists of the remaining psychiatric ratings. The psychiatrists appear to perceive these items as less job related than the others and often do not even score them. Factors three and four are based on the psychologist's ratings. Factor three may be conceptualized as a dimension concerned with evaluation a candidate's fit or match with the police work and this particular police organization. Factor four, on the other hand, appears to be a more general dimension related to overall achievement.

F. Final Employment Status. Each candidate's final employment status, accepted or rejected, was regressed against 1) the police department's evaluation, 2) the psychologist's evaluation, 3) race, 4) sex, and 5) age, again, stepwise linear logistic regression was used. Of the variables, only race did not contribute to final employment status. Follow-up calculations, using the logistic regression coefficients, indicated the following relationships between the variables and final employment status. For male candidates, the probability of being selected was 70%. The probability for female candidates was 88%. In other words, women were more likely to be accepted than men. Applicants who were one standard deviation older than the mean of the sample (about 28 years old), had a 72% probability of being accepted. In contrast, applicants had a standard deviation below the mean, 20 years old, had an 87% probability of being accepted. If a candidate was acceptable to police based on the background investigation and polygraph results, the probability of that candidate being accepted at the final decision point was 95%. However, if the background investigation and polygraph results were unacceptable to police, the candidate's probability of being accepted fell to 50%. Likewise, if the psychologist found a candidate suitable for police work, the candidate's probability of being accepted for employment was 92%. If the psychologist recommended further evaluation, the candidate's probability of being accepted for employment fell to 61%.

FIGURE I ILLUSTRATION OF THE POLICE RECRUIT HIRING PROCESS.



* * *

IPMAAC STUDENT PAPER AWARD

Chair: Thomas E. Cressler, Tennessee Valley Authority

The Influence of Sex Stereotyping and the Sex of the Job Evaluator on Job Evaluation Ratings

Anne Marie Carlisi, Bell Communications Research,
Livingston, New Jersey

Organizations have traditionally used job evaluation to establish systematic wage and salary structures. The worth of job evaluation procedure is established to the degree that jobs possess certain requirements or characteristics, such as skill, effort and responsibility requirements, and the conditions under which the job is performed. In theory, then, job evaluation should provide a systematic means of comparing similarities and differences among jobs according to their relative contribution or value to the organization, for the purpose of setting equitable wage and salary rates.

Despite the intended purpose of job evaluation, comparable worth theorists seem to be ambivalent about its use in the wage and salary setting process. They concede its potential as a bias-free method for determining the comparable worth of jobs within organizations (Treiman & Hartmann, 1981). In fact, Remick (1984) has operationally defined comparable worth as "the application of a single, bias-free point factor job evaluation system within a given establishment, across job families, both to rank-order jobs and to set salaries." At the same time, however, the advocates as well as the opponents of comparable worth have questioned whether biases that may occur during the process of job evaluation will result in devaluation and lower pay rates for traditionally female jobs.

The present study addressed two of the potential biases cited by the critics of job evaluation. First, the critics have questioned the reliability of job evaluation proposing that the subjectivity of the judgements involved in making job evaluation ratings allows for the possibility that male and female raters evaluate jobs differently. The influence of rater's sex became an issue in job evaluation following a study by Arvey, Passino and Lounsbury (1977). They found that female job analysts gave consistently lower scores than did males to jobs on Position Analysis Questionnaire (PAQ) dimensions, which is used as a job evaluation instrument (McCormick, Jenneret, & Mecham, 1972).

The second potential bias examined in the present study was the influence of sex stereotyping which may cause job evaluators to unintentionally devalue jobs typically performed by women. Specifically, job evaluators may cognitively simplify the rating task by categorizing jobs as masculine or feminine. Therefore, job evaluators may subsequently make job evaluation ratings based on catagory-relevant information (i.e, the sex stereotype of the job) when there is an insufficient amount of job information or in lieu of having to process large amounts of information about each job.

A substantial amount of research in the area of sex stereotyping (for a review see, Ashmore & DelBoca, 1981) reveals that this categorization process results in sex bias. Specifically, stereotypically female characteristics, performance, and occupations are perceived as less valuable than those of the stereotypical male, If job evaluation ratings are similarly sex-biased, and are used to determine the comparable worth of jobs, female job of comparable worth to a male job may indeed be undervalued.

With these criticisms in mind, the present study was designed to examine whether the sex of the job evaluator influenced job evaluation ratings, whether commonly held sex stereotypes led to lower ratings of traditionally female jobs in comparison to traditionally male jobs, and whether the amount of information presented to job evaluators moderated the effects of sex stereotyping on job evaluation ratings. It was hypothesized there would be no differences in job evaluation ratings made by males and females. It was also hypothesized that as information about the jobs increased, the effects of sex stereotyping would decrease and male and female jobs of equal worth would receive similar job information, female jobs would be rated lower in more limited information conditions than in information conditions containing substantial amounts of specific job information.

Method

There were 308 participants in the study. One hundred fifty-four male and 154 female graduate and upper level graduate students served as job evaluators. Subjects evaluated four jobs: drafter, legal secretary, shipping and receiving clerk, and file clerk. These jobs were selected in order to obtain two pairs of jobs of equal worth. Worth, in this study, was operationally defined by job prestige. Prestige ratings were used because they have been found to correlate highly with both income and worth as measured by job evaluation points. Prestige scores for each job were obtained from Treiman's Occupational Prestige Scale (1977). The jobs of drafter and legal secretary were identical worth. Both jobs were of medium prestige. The jobs of shipping and receiving clerk and file clerk were low prestige jobs.

In each pair of jobs, one job was stereotypically male and the other stereotypically female. In choosing jobs this way, any differences in the job evaluation rating of the two jobs in each pair could confidently be attributed to the stereotype of the job, rather than to actual differences in job worth. The stereotype of jobs was determined on the basis of the percentage of males and females employed full time in each occupation as specified by the Statistical Abstracts of the United States (1980). A 75% cutoff for participation rates was used.

The job information on which the job evaluation ratings were based was either in the form of a job title, a job description, job specifications, or all combinations of these three. In all, there were seven information conditions.

The job evaluation instrument was the Comprehensive Job Evaluation Technique (CJET). The CJET is a point method of job evaluation. The instrument consists of four main factors: Skill, Effort, Responsibility, and Working Conditions. In all, there are 15 scales in the instrument. Skill is broken down into three scales: Education, Previous Experience, and Time to Proficiency. Effort consists of four scales: Mental Effort, Visual Attention, Physical Effort, and Manual Dexterity. Responsibility consists of five scales: Supervisory Responsibility, Financial Responsibility, Responsibility for the Safety of Others, Counseling and Teaching, and Negotiating and Influencing. Working conditions are broken down into three scales: Surrounding, Hazards, and Monotony. For each scale the possible scores range from one to five.

The primary statistical analysis was a 2 x 2 x 2 x 7 repeated measures analysis of variance. The dependent variables were total job evaluation points for each job. The manipulated variables were job stereotype and job prestige (within subjects variables), and sex of rater and information (between subjects variables).

Results and Discussion

The analysis revealed no significant differences in job evaluation ratings made by male and female raters. There was, however, a consistent sex bias toward stereotypically female jobs, which were rated significantly lower than male jobs of equal worth in all information conditions (see Figure 1).

Contrary to what was hypothesized, female jobs were not undervalued more in limited information conditions (job title only, job description only, and job title-job description). It was expected that in these limited information conditions, raters would be forced to rely more on job stereotypes than in conditions where job specifications were included. Job specifications were believed to provide raters with very specific, job-relevant information on which to base their job evaluation ratings. This type of diagnostic information has been found to eradicate the effects of sex stereotyping in previous research. The opposite was found in this study. Curiously, the inclusion of job specificaitons did not diminish the effects of stereotyping on job evaluation ratings. In fact, the largest mean difference between the ratings of male and female jobs occurred in the information condition in which raters were provided with just job specifications, and no job title to identify the jobs.

Contrary to past research, diagnostic job-relevant information was ignored by the job evaluators in this study, and gender cues were used to infer stereotypic job characteristics. Specifically, the job information presented in the job specifications was identical for the male and female jobs in each pair except for one word in each set of job specifications (i.e., drafting-typing, warehouse-office). It appears that these words served as gender cues which influenced job evaluation ratings more than specific job information. Therefore, when raters were given identical information about the performance requirements for male and female jobs, and were not given the job titles as primes for the use of sex stereotypes, they still produced ratings strongly reflective of commonly held stereotypes.

An alternative explanation for these findings in the job specifications only information condition is that the job evaluation ratings may be reflective of perceived skill differences, especially in the medium prestige pair of jobs. Specifically, with no title to identify the jobs, the words drafting and typing may have led raters to believe the jobs were engineer and typist. In this case, the ratings may be more indicative of the perceived differences in skill level rather than sex stereotypes.

If the devaluation of the female job were due to skill differences, however, there should be larger differences between the ratings of the medium prestige, male and female jobs, than between those of low prestige. However, the results reveal no significant job prestige X job stereotype interaction in the job specifications condition. The presence of significant main effects for job prestige and job stereotype, but no significant interaction for the two, indicates that perceived skill did not result in the differences in the job evaluation ratings.

Overall, the results of the present study provide conflicting practical implications for the users of job evaluation instruments. First, it was shown that the speculation of rater sex bias is unfounded. This provides encouraging results for the continued and increasing use of job evaluation systems by organizations. Second, the speculation that sex stereotypes bias the evaluation of female jobs may, indeed, be the truth. The results of the present study provide preliminary evidence that female jobs tend to be devalued with respect to male jobs of comparable worth.

REFERENCES

Arvey, R.D., Passino, E.M., & Lounsbury, J.W. (1977). Job analysis results as influenced by sex of incumbent and sex analyst. Journal of Applied Psychology, 62, 411-416.

Ashmore, R.D., & DelBoca, F.K. (1979). Sex stereotypes and implicit personality theory: Toward a cognitive-social psychological conceptulization. Sex Roles, 5, 219-248.

Doverspike, D., Carlisi, A.M., Barrett, G.V., & Alexander, R.A. (1983). Generalizability analysis of a point method job evaluation instrument. Journal of Applied Psychology, 68, 347-368.

McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 56, 347-368.

Remick, H. (Ed) (1984). Comparable Worth and Wage Discrimination: Technical Possibilities and Political Realities. Philadelphia: Temple University Press.

Schwab, D.P. & Grams, R. (in press). Sex related errors in job evaluation: A "real-world" test. Journal of Applied Psychology.

Treiman, D.J. (1979). Job evaluation: An analytic review. Washington, D.C.: National Academy of Sciences.

Treiman, D.J. & Hartmann, H.I. (Eds.). (1981). Women, work, and wages: Equal pay for jobs of equal value. Washington, D.C.: National Academy Press.

U.S. Bureau of the Census. (1980). Statistical abstracts of the United States. Washington, D.C.: U.S. Government Printing Office.

Figure 1

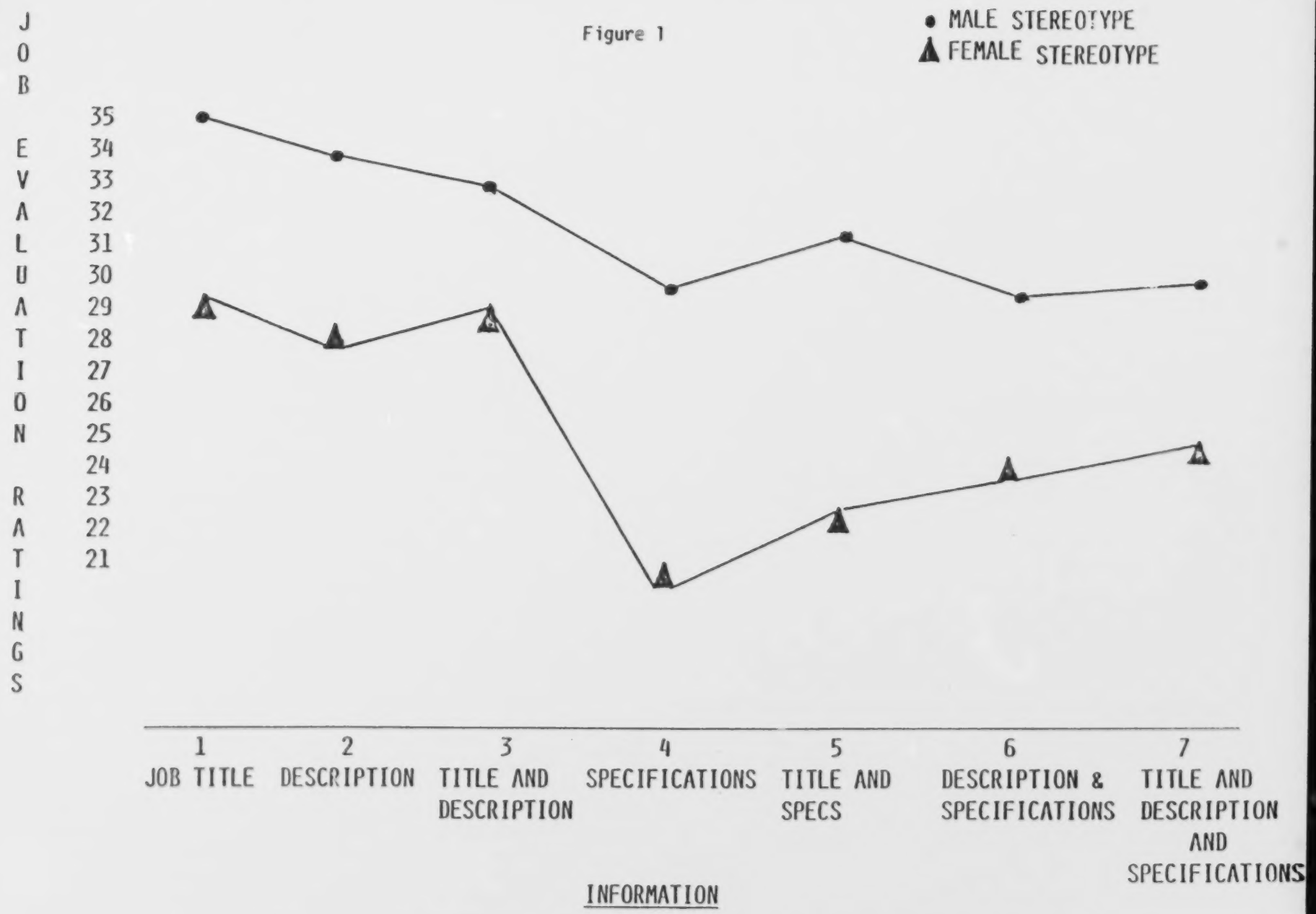


FIGURE 1: THE INTERACTION OF INFORMATION AND JOB STEREOTYPE IN DETERMINING MEAN JOB EVALUATION RATINGS.

IPMAAC SPECIAL SESSION: Equitable Compensation: Methodological Criteria for Comparable Worth

Moderator: James C. Johnson, State of Tennessee

Discussants: Thomas A. Mahoney, Vanderbilt University and
Francis S. Guess, Member, U.S. Civil Rights Commission

Technical Standards for Comparable Worth Implementation

Ronnie J. Steinberg, Senior Research Associate, Center for Women in Government; Associate Professor, Public Affairs and Policy and Sociology, State University of New York at Albany

Comparable worth concerns the issue of whether work done primarily by women and minorities is systematically undervalued because the work has been and continues to be done primarily by women and minorities. By systematic undervaluation, we mean that the wages paid to women and men engaged in historically female or minority work are artificially depressed relative to what those wages would be if these jobs had been and were being performed by white males. Simply stated, comparable worth involves correcting the practice of paying women less than men for work that requires equivalent skills, responsibilities, stresses, personal contacts and working conditions.

Over the last decade, the policy has evolved to correct the sex and race discrimination in wages that results from occupational segregation. The link between segregation and the wage gap is now undeniable. In 1981, the National Research Council of the National Academy of Sciences (NRC/NAS) concluded, on the basis of three years of analysis, in their final report Women, Work and Wages, "Not only do women do different work than men, but the work women do is paid less, and the more an occupation is dominated by women the less it pays" (Treiman and Hartmenn, 1981: 28). And again, they wrote: "Women are systematically underpaid . . . on the basis of the review of the evidence, our judgement is that there is substantial discrimination in pay" (Ibid., 66-67).

The wage gap between women and men is one of the oldest and more persistent symptoms of sexual inequality in the United States.

> As of today, female workers employed full-time year around earned, on average, around 64¢ to 65¢ to an average male worker's dollar. This is similar to figures for 1955. In 1974, the wage gap dropped to 57¢ on the dollar.

If we break down this wage gap figure by race, the statistics are even more disturbing. As of 1982, black men earned 76¢ relative to the $1.00 earned by men on average; black women earned 56¢ relative to this dollar standard; and hispanic women earned 52¢ on the dollar.

In addition, to the fact that there is a wage gap, is the fact that women are concentrated in a narrow range of low-paying occupations. In 1982 more than 50% of all female employees are found in only 20 of a total of 427 occupations. These 20 occupations are among the lowest paid clerical and service jobs. Moreover, study after study has found that the single most important source of the wage gap between women and men is occupational segregation.

Not all of the observed gap in wages is due to discrimination, however. Occupational segregation can translate into wage differences between women and men for two reasons: first, women may be segregated into jobs that require less skill, effort and responsibility than jobs filled by men. In other words, there are real differences between jobs held by women and jobs held by men. This is an affirmative action issue but not a comparable worth issue. Second, women may be segregated into lower paying jobs that require the equivalent amount of skill, effort, and responsibility as male jobs. This latter difference in wages that cannot be accounted for by differences in the value (to the employer) of the work performed is what is meant by systematic undervaluation of work or wage discrimination.

Because pay equity addresses wage discrimination that is a by-product of occupational segregation, it is necessary to understand what policies and practices reinforce and perpetuate the situation that female-dominated or significantly-minority work is not compensated at an equivalent rate with jobs performed by white males. In the area of compensation, the institutional policies under scrutiny are classification systems, a majority of which are built out of some variant of job content analysis and job evaluation. Job content analysis and evaluation methodologies need not lead to sex- and race-based wage discrimination, however. Job evaluation is a technique for making systematic and explicit the values operating in a specific labor market. These values are described in terms of what people do on their jobs. Job evaluation also provides a procedure for systematically ordering jobs into a relative wage structure based upon the values articulated. In practice, however, the way job evaluations have been designed and carried out in most firms have had the effect of creating a two-tiered pay policy, in which sex- and race-type of job are implicit job content characteristics that operate to depress the rate of pay. This should not be surprising.

New York State provides a characteristic illustration of the way in which many existing systems embed wage discrimination in the description and evaluation of jobs. New York State uses a job evaluation system which groups particular positions into job classes like Secretary I, Secretary II, Cook, or Carpenter. Classes are then assigned to one of 85 job families or occupational groups, such as tax administrators and technicians, parks and forestry, general clerical and food preparation. Within each occupational group, classes are arranged hierarchically from highest to lowest in terms of job content characteristics. No points are assigned to the characteristics. Each occupational group is then attached to a general grading scheme independently. This means that there is no comparison with other occupational groups that may have similar job content characteristics.

Because there is no internal equity across job families, the New York State system carries 85 metrics or standards of worth--one for each occupational group. This way of designing job evaluation has been labeled the multiple plan problem by Hartmann and Treiman (1983), the authors of the National Academy of Science study. It occurs when major sex- and race-segregated occupational categories like clerical jobs, manual labor jobs, and managerial jobs are treated independently and differently from each other. Descriptions are frequently based on dissimilar job content features. Evaluations are based on different factors. Similar factors used in different occupational groups are given different weightings. Salaries are set in relation to different external firms. It is troublesome for comparable worth because it prohibits comparisons across categories.

A second way in which cultural assumptions embed wage discrimination into classification systems involves the differential description of jobs. This occurs when compensable job content characteristics of female-dominated and significantly minority jobs are not gathered or are overlooked or ignored.

This first example of this shortcoming is drawn from a University of Wisconsin extension school study of the 3rd edition of the Dictionary of Occupational Titles (DOT) (Witt and Maherny, 1975). The DOT, compiled by the U.S. Department of Labor, contains a list of almost every job title along with a rating of the job in terms of a skill-complexity code. The skill-complexity code is built on the assumption that "every job requires a worker to function at some definable level with regard to Data, People and Things" (Ibid. 24). These researchers were disturbed by the ratings given to certain types of predominantly female jobs compared to certain predominantly male jobs. For instance, dog pound attendant and zoo keeper were rated more highly than nursery school teacher or day care worker. The researchers carried out an independent assessment of the predominantly female jobs. Their ratings differed substantially from those of the Labor Department evaluators.

When examining why the differences emerged, they found that the Labor Department evaluators had overlooked important characteristics of the female-dominated jobs, especially those associated with taking care of children. The evaluators did not regard these as job related skills but rather as qualities intrinsic to being a woman. In other words, the job evaluators were confusing the content and responsibilities of a paid job with stereotypic notions about the characteristics of the job-holder. This is often done with respect to fine motor coordination and rapid finger dexterity in female-dominated blue collar and clerical work.

A second example is taken from a job evaluation manual comparing the rating of experience for typist and truck driver (Treiman, 1979: 52-53). To score the job knowledge factor of this system, it was necessary to determine how much pre-job and on-the-job experience would be needed to perform the job duties under normal supervision. The typist was judged to require one month of training time and truck driver was judged to require twelve months training time. There was no discussion as to why it was judged that truck driving required 12 times as much training as typing.

On this job content characteristic alone, the truck driver's salary was two pay grades higher than the typist's.

A final example is drawn from an examination of sixteen job analysis/job evaluation frameworks I conducted as a preliminary step in developing a customized job content questionnaire for the New York State comparable pay study (Steinberg, Possin, Treiman, 1984). We reviewed other existing evaluation schedules to include, in our questionnaire, every category of job content characteristic someone had found to be compensable. We also built in the levels or degrees appropriate to predict compensation that had been used in other systems. This one hundred page item list proved incomplete as we began to read over job specifications in the major New York State job families and to conduct preliminary field-testing of the questionnaire with incumbents of key jobs. Previous approaches had either overlooked certain characteristics associated with female- and minority-dominated work or else had formulated questions in such a way that people in institutional and facility human service settings (largely women and minorities) would have read as not applicable for them to answer. In putting together the Job Content Questionnaire for the New York State Comparable Pay Study, we developed a preliminary checklist of frequently overlooked job content characteristics found in female-dominated jobs.

In this and the last set of examples, wage discrimination would be a function of the fact that the prerequisites and tasks of jobs historically filled by women and minorities have been ignored, forgotten, overlooked, or regarded as unnecessary of compensation. The source of this oversight is, again, primarily cultural. Comparable worth job evaluation studies seek to remove these and other discriminatory components operating in current salary setting procedures.

Specifically, we believe that for a job evaluation study to have incorporated gender equity concerns into its research design, it must meet the following criteria, which I have organized in terms of three job evaluation components -- description, evaluation, and salary setting.

(1) Description: All jobs must be described fully and consistently and not differentially by the sex or race of the typical incumbent. This means that jobs must be viewed in terms of the same possible range of job content characteristics. These characteristics must include ones associated with female-dominated or significantly minority work.

(2) Evaluation: All jobs must be evaluated and assigned points according to a uniform set of factors and factor weights. It does not matter whether the factors are obtained from a a priori standardized system, from an a priori customized model in which policy-makers generate a set of factors and weights from scratch, or from a policy-capturing model. Factors must encompass characteristics associated with female-dominated and significantly minority work. The evaluation framework must be applied consistently across all titles.

(3) Salary-setting: All jobs in a firm should be assigned wages according to one pay policy line. However, this line must be adjusted for possible discrimination in market rates using one of three possible adjustment formulas. If jobs are benchmarked to the external labor market, both firms and job titles must be comprehensive and representative of the labor markets involved.[1]

We tried to meet these standards in designing the New York State pay equity study. In other words, our goal was to maximize consistency and minimize sex and race bias in the way jobs are described and evaluated and in the procedures for establishing wages.

The size of the New York State employment system coupled with time and money limitations increased the challenge in meeting these objectives. Currently, the state system encompasses over 6,000 job titles affecting over 170,000 employees, almost 50 percent of whom are women and 22 percent minorities. The classification and compensation system was established in 1937 and last revised in the 1950s. It has never been assessed to determine the effects of sex and race segregation on the setting of salaries.

The questionnaire used in the New York State pay and equity study was designed to include questions about job content that would predict the current wage structure, include questions about job content that would be found in female-dominated and significantly minority jobs and include questions that would allow for comparisons across job titles differing by sex and race. The questionnaire currently contains 110 specific items (Exhibit E). For each question, employees must choose one from among a number of possible responses provided to them. All responses are closed-ended. In this way, we will be asking the same question to employees in many different job titles.

The set of factors and factor weights will be developed using policy-capturing evaluation. In New York State, this means that they will be derived directly from the data collected from state employees through a self-administered questionnaire. This eliminates the possibility that consultants or evaluation committees impose stereotypes as ambiguous job descriptions. To be sure, employees carry these stereotypes as well. Yet, we have asked specific and factual questions about jobs and we have asked the same questions to all employees. In addition, our procedure involves averaging incumbent responses to obtain a composite job description. In some cases, we will be averaging the responses of 50 employees within one job title. This averaging process, combined with a detailed questionnaire, provides, to our knowledge, the best available methodology for accurately capturing job content information on an employee population of this size.

Since we are interested in examining the job title, and not the individual incumbent, sampling considerations were complicated. Our sample was not hand-picked by either labor or management. Rather, we drew a representative sample of all New York State job titles using

systematic sampling procedures. Specifically, we sampled all job titles in grades three to 22 with four or more incumbents and all job titles in grades 23 to 38. This enabled us to include single incumbency management positions in our sample. The approach gave us approximately 2900 job titles in our sample.

To sample individual incumbents within each title, we drew two different samples, one for targeted female-dominated and significantly minority titles and a second for all other titles. For the targeted titles, we gathered information from all incumbents in titles with 150 or fewer incumbents. In titles with more than 150 incumbents, 150 incumbents were sampled. For the other titles, we gathered information from all incumbents in titles with 20 or fewer incumbents. Where there were more than 20 incumbents, we sampled 20 incumbents. The sample we developed to maximize representativeness in the range of job titles and minimize the error of the estimates of wage discrimination we would make by gathering as much information from as many incumbents as possible.

The final sample included 37,087 employees working throughout New York State. The survey was distributed only through a mailed survey, with two follow-up letters between December 1984 and February 1985. The survey response rate was 73 percent.

We are in the process of analyzing the data statistically. Thus far, we have taken the information from the Job Content Questionnaire and grouped them to create factors. We have uncovered 14 factors that capture job content in New York State. They include: management/ supervision; working conditions; communications with the public; computers; stress; fiscal responsibility; autonomy; group facilitation; and information use. Not all of these 14 factors will predict pay. How many and which ones do predict pay for New York State is what we are examining now.

We are doing this by developing a policy-capturing model for the New York State government employment system. This means statistically establishing the relationship between the current wages paid for in the New York State system and the content of these jobs.

Once we have established the model for the State system as a whole, we will statistically adjust it using two procedures proposed by the National Acedemy of Sciences study committee. One procedure involves using the white male pay policy line as a standard against which to predict the pay for all jobs. A second involves adjusting the overall compensation model by statistically removing the effects of percentage female and percentage minority on the job content characteristics predicting pay. It is important to adjust this overall pay line, because it includes, as part of the average, the current wages for female-dominated and significantly minority jobs, in which discrimination may be embedded. Without adjustment, the overall pay policy line could even result in embedding discrimination into the wages of male jobs set in relation to it and lowering those wages.

Once we have obtained the pay equity estimates, we will report the results to the Civil Service Employees Association, AFSCME and the Governor's

Office of Employee Relations, co-sponsors of the project. Selection of the final adjustment equation on which equity estimates will be based will be made by labor and management, as will the procedures to follow in implementating pay equity adjustments.

---

[1]Sections of this paper are drawn from Steinberg, 1984: Steinberg and Haignere, 1984a; Steinberg and Haignere, 1984b; and Steinberg and Haignere, 1985.

BIBLIOGRAPHY

Hartmann, Heidi I., and Donald J. Treiman. 1983. "Notes on the NAS Study of Equal Pay for Jobs of Equal Value," Public Personnel Management 12 (Winter): 404-417.

Schwab, Donald P. 1980. "Job Evaluation and Pay Setting: Concepts and Practices," in Comparable Worth: Issues and Alternatives, edited by E. Robert Livernash. Washington, D. C.: Equal Employment Advisory Council.

Shepela, Sharon Toffey, and Ann T. Viviano. 1984. "Some Psychological Factors Affecting Job Segregation and Wages," in Comparable Worth and Wage Discrimination, edited by Helen Remick. Philadelphia: Temple University Press.

Steinberg, Ronnie. 1984. "Job Evaluation Methodologies and Comparable Worth Policy," paper prepared for a seminar on New Concepts and Research Directions in Pay Determination, Cornell University, November 8-9.

Steinberg, Ronnie and Lois Haignere. 1984a. "Separate But Equivalent: Equal Pay for Work of Comparable Worth," in Gender at Work: Perspectives on Occupational Segregation and Comparable Worth. Washington, D. C.: Women's Research and Education Institute.

Steinberg, Ronnie and Lois Haignere. 1984b. "Review of Massachusetts Statewide Classification and Compensation System for Achieving Comparable Worth," unpublished Report submitted to the Commonwealth of Massachusetts Special Committee on Comparable Worth. Albany, NY: Center for Women in Government.

Steinberg, Ronnie and Lois Haignere. 1985. "Equitable Compensation: Methodological Criteria for Comparable Worth," paper prepared for a conference, Ingredients for Women's Employment Policies, State University of New York at Albany: April 19-20.

Steinberg, Ronnie, Carol Possin, Donald Treiman. 1984. Job Content Questionnaire. Albany: Center for Women in Government.

Treiman, Donald J. 1979. Job Evaluation: An Analytic Review. Washington, D. C.: National Academy of Sciences.

Treiman, Donald J., Heidi I. Hartmann, and Patricia A. Roos. 1984. "Assessing Pay Discrimination Using National Data," in Comparable Worth and Wage Discrimination, edited by Helen Remick. Philadelphia: Temple University Press.

Treiman, Donald J., and Heidi I. Hartmann. 1981. Women, Work and Wages: Equal Pay for Jobs of Equal Value. Washington, D. C.: National Academy Press.

Witt, Mary, and Patricia K. Nahemy. 1975. Women's Work--Up from 878: Report on the DOT Research Project. Madison, Wisconsin: Women's Education Resources, University of Wisconsin-Extension.

* * *

COMPARABLE WORTH (Paper Session)

Chair: Linda Davey, South Carolina Division of Human Resource Management

Discussant: Dennis Doverspike, University of Akron

The New Frontier for Public Human Resource Management

Claire J. Anderson
Loyola University in New Orleans

Introduction

The issue of comparable worth entails equal pay for jobs of comparable value. Comparable worth is no longer an obscure legal issue nor is it any longer the cause celebre of the vocal few. Rather, it is practically a household word given its wide publicity and debate. Today, the general tendency is to focus on the legal aspects and to debate the pros and cons of the applicability of Equal Employment laws or needed legislation. The environment of comparable worth reflects the trend of the times -- to look to legislative and judicial institutions to define issues, solve problems and enforce decisions insofar as civil rights are concerned. Thus far the US Supreme Court is not ready or willing to rule on comparable worth, although, the growing number of lower court decisions and legislative action on the part of the states bring home clearly that comparable worth is following the same path as other human rights issues in the workplace, and that there is more and more dependence on legal action and government intervention rather than one of initiative on the part of public and private employers. Historically, inequity in the work place has been mandated by government--the arena wherein a large but relatively weak population segment can find redness of inequities. The list is long and includes practically every major social change in employment in the 20th Century: child labor laws, prohibition of bare subsistence wages, protection for collective bargaining and equal employment opportunity. The record is clear that government has intervened where free enterprise has failed to do so. Despite advances, today it is still a fact that even with all other things equal men still enjoy wages well in excess of that of women.

The premise of this paper is that Human Resources Managers can be the keystone in meeting the comparable worth controversy by addressing the problem at its grassroots rather than perpetuating the inequities of the past or looking to government for action. Focus is on the comparable worth of dissimilar rather than similar jobs.

A Framework for Analysis.

For comparable worth to attain any foothold, a radical change process is inevitable. For the purposes of predicting change, a general model of the change process was adopted. A society (or organization) at any

given time is a dynamic balance of forces supporting and restraining any practice. While the system is in a state of relative equilibrium so that current practices will continue in a steady way until change is introduced. Change will occur by increasing supporting forces for it and/or reducing the restraining forces. As far as comparable worth is concerned, the restraining and supporting forces span a number of psychological, sociological, ethical, economic and business issues all of which are supplementary rather than mutually exclusive. This paper will briefly address the nature and strength of these forces and the potential role for Human Resources as a major force in change.

The Social Roots of Wage Differentials

The idea of comparable worth while not the first "women's idea" is probably the purest. Despite gains in civil rights, wage gap is the most persistent symptom of sex inequality as women still predominate in a few occupations, all of which are low paid. One of the best known statistics is that on an average a woman makes a salary somewhere around 60% that of the average male. What is lesser known is that this gap has not changed significantly since the 1940's. This is hardly surprising as historically and without exception women's work is valued below that of a man whether it be done in the household or the workplace.

Reasons for the differences are put forward from a number of viewpoints with the most common being that the career patterns of males in the past were relatively uninterrupted except probably by a few years of military service whereas the career pattern of women rarely went beyond that of the birth of the first child. Thus, women had less experience, less training and less attachment to the workplace and therefore should be paid less. Some even argue that women's wages <u>should</u> be lower than men's because most men are primary breadwinners; but even if this were a viable argument, it ignores the trend of the times with the exponential growth of the single parent family where women are by far the primary wage earners. The second apologia centers on the fact that occupations to this day are gender segregated with the vast majority of women concentrated in a very few occupational categories which incidentially are low paying. Thus as long as women tend to cluster in lower paying jobs naturally earnings will be less. This argument however fails to recognize the "critical mass" phenomenon; that is, as women enter any one occupational field in numbers the pay decreases. The more women entering a field, the less likely they are to receive the same wage as that of their male counterparts (14). One analysis found that for any one percent change of women in any occupational grouping there will be an average decrease of $42 in annual earnings (26). In the 19th Century secretaries were highly regarded and well paid positions for a young man wishing to make entry into the business world. With the invention of the typewriter, believed to be well suited for girls, the job of secretary became a woman's job and the pay dropped (29). The entry of males into a field also appears to raise salaries but thus far the impact is minimal if for no other reason than men have yet to enter traditional female occupations in any numbers.

Proponents hold that much of the wage difference is rooted in historical social norms which undervalue "women's work". Take for example the question of the university professor. A man who teaches part time as an adjunct to another job is highly valued whereas a woman whose part time teaching supplements home duties is far less valued. In a far more visible case, consider that despite enlightened approaches to child care and lip service given to the importance of early childhood in a person's future well being and cognitive development, and the burgeoning demand for quality child care, workers in this field (almost exclusively female) are notoriously underpaid. This pervades at all levels. One author recalls the summer employment experiences of her children. The son took a job mowing lawns and made the equivalent of $8 per hour which allowed him to spend his afternoons on the beach. His sister took a job caring for children, made meals, worked from 8 till 5 at approximately $2 per hour. Does society value its lawns more than its children?

While it is not particularly in vogue in the middle 1980's, further insight into the wage gap may be gained in feminist writings. Twenty years after Betty Friedan published The Feminine Mystique, she again opened the door to a new feminist development in The Second Stage. She holds that at least in the past, men and women have thought and acted differently especially when called upon to be leaders. She terms these "Alpha" and "Beta" modes (rather than masculine and feminine but the message is clear):

> "Alpha-style leadership ... is based on analytical rational thinking. It relies on hierarchical relations of authority ... Alpha is more 'direct' and 'aggresive' ... strives competitively for all-or-nothing solution, expecting a 'clear win-or-lose' ... with 'any non-win conclusion resulting in a loss of face.'"
>
> On the other hand, the Beta or feminine style is "based on synthesizing, intuitive, qualitative thinking.... It is tuned to more complex, more open and less defined aspects of reality. Its concern is the whole picture being presented rather than fixed quantities and the status quo".(11)

Friedan points out that the division is not biological but rather that women perfected the Beta mode because their province was family life, whereas men perfected the Alpha mode because of the work they did. This is closely allied to Carol Gilligan's book In a Different Voice which furthers this idea, concluding that women and men have different moral domains:

> Women's construction of the moral problem as a problem of care and responsibility in relationships rather than as one of rights and rules ties the development of their moral thinking to changes in their understanding of responsibility and relationships, just as the conception of morality as justice ties development to the logic of equality and reciprocity. Thus the logic underlying an ethic of career is a psychological logic of relationships which contrasts with the formal logic of fairness that informs the justice approach.

The issue here is that of the "maleness" of the workplace and the market. That is, that employment is competitive and pleasant, lucrative work is scarce, employers do not necessarily reward virtues of "caring" and "responsibility for others" and proficienty in behavior now labeled "feminine" does not guarantee a person a job of their choice. The dilemma of women entering the work place is one either resigning themselves to remain at the bottom of the career ladder or giving up their "traditional feminine qualities". Not all advocates of comparable worth repudiate the "male mode" and criteria traditionally used to evaluate men's jobs. The example cited is that of the profession of nursing--which incidentally gave rise to the comparable worth issue--wherein the virtues of nurturing and caring predominate. Feminist thought puts forth the question: "Must women act and think like men to gain economic parity"?

Ethical Considerations

The comparable worth controversy also entails the problem of whether or not we can morally justify the "free market" principle in wage setting. The question here is how does comparable worth advance the cause of individual rights and dignity, how does it promote justice or a fair distribution of society's goods and how does it contribute to the productive or efficient use of social resources. This may be approached through examination of two theories of ethics: The rule (deontological) theory or a principled commitment to some fundamental concept, e.g. truthfulness regardless of consequences, the "Golden Rule" as opposed to a "results" (teleological or consequential) theory which focuses on results or consequences" or the ends justify the means. If, for example, truthfulness rather than falsehood is perceived as a means of better serving the public good then be truthful; but, do not commit yourself to the principle of truthfulness because there may be times when falsehood will serve better.

The question then arises as to whose interests should be preserved. No one seriously believes that a firm should jeopardize its competitive position by unilaterally adopting the comparable worth concept, however, many firms go beyond the minimum of law. Can the cause of justice and efficiency be advanced simultaneously? If not, which is in the best interest of the public good? Such questions are not easily answered. A further ethical consideration involves the educational community, particularly higher education and professional organizations which place formidable formal education barriers into certain professions (nursing, librarianship to name a few) many of which are historically female dominated. Human capital theory and common sense would predict that long years of preparation for any profession raises expectations of at least intrinsic rewards. Frequently when the reality of work life sets in and intrisic rewards are not realized, the search for extrinsic rewards--viz. money, sets in and, when this does not materialize, even deeper disappointment and resentment sets in.

Economic Considerations

In writing the opinion in the now famed Lemons case, Chief Judge Fred Winner, of the U.S. District Court of Colorado in familiar legal jargon stated that the concept of comparable worth was "pregnant with the possibility of disrupting the entire economic system of the United States of America." Portents of major disaster in the free market economy are not new. Economic analysis fails because of a number of fallacies: the theory of the "free market"; treating humans in the same manner as any other commodity; failure to recognize the nature of open and closed labor markets and in estimates of inflation. Neoclassical theory describes a concept wherein purchasers and sellers of labor compete in a market where supply and demand are adjusted through the price mechanism with a prevailing price being the "market clearing price." Employers and workers engage as equals in bargaining for wages on the relative supply and demand for workers within a given qualification. Thus, workers will seek out those offering maximum wages for services the workers have to offer and employers will seek to minimize the wage bill while maintaining the desired quality of the workplace.

The failure of neoclassical theory is manifest daily with chronic shortages in clerical workers but the field remains relatively low paid. Clerical work is "woman's work" and many women opt to drop out of the labor market or seek alternative employment rather than work for "peanuts." Human capital theory holds that investments in people yield monetary payoffs by making labor more productive. A person's human capital appreciates through things such as schooling or job experience. This then explains that because women's employment is intermitten due to domestic responsibility, their skills depreciate while they are outside the workforce rather than appreciating as they would be in any job. Thus women who anticipate intermittent employment will choose occupations that require skills that do not account for discriminatory practices such as the case of a Wharton female MBA who was told by a placement director that generally each year of work experience after college is worth $1,000 in salary but the six years as a nursing supervisor would not be considered worth anything because it was a woman's field. Ex-teachers, social workers, librarians and others have the same problems whereas male classmates who by their own admission spend time in the military "smoking grass and goofing off" got higher starting salaries.

Clarence Pendleton of the US Civil Rights Commission noted for his "Looney Tunes" description of comparable worth commented on the impact of comparable worth on the free market as follows: "I think you just cannot begin to do things to the marketplace that have served this country so well" but one must remember that this same free marketplace did not eliminate slavery, child labor or discrimination in employment. The final economic argument is that if wages go up employers will resort to automation thus fewer jobs will be available--a prediction frequently made but rarely if ever fulfilled. Economic views of efforts at pay equalization under comparable worth generally fall under the rubric of "social engineering" and an attempt to interfere with the economic system which will ultimately result in social and economic disaster.

The Human Resource Role- Job Evaluation

Proponents of immediate remedy to comparable worth problems place a good deal of faith in the job evaluation systems which purport to measure the "worth" or "value" of a job (not a person) to the employer. Job evaluation systems are not new and in fact were used extensively by the Labor Powers Board during World War II. Opponents of comparable worth attack job evaluation systems using the old "apples and oranges" comparison argument; however, even the allegory does not hold as apples and oranges can be compared for example in terms of calories, sweetness, minerals, vitamins, etc.

Job evaluation systems are far from scientific and fall far short of validity and reliability. They generally use some quantitative measure (points, weights) thus give an aura of precision which is questionable at best. Such systems are assumed by many as "gender free" but such systems have been found to assign extra weight to characteristics such as physical strength in which women are unable to excel and ignore items such as motor control and rapid movement with a low error rate--a feat more common among women. Job evaluation has been attacked as "the single most effective device by which organizations retain and create discriminatory pay practices."

Conclusion

No adequate functioning model of comparable worth exists today; although, some 20 states have enacted statutes that contain either comparable worth or comparable character language. However, statutes vary immensely. Some simply mandate review of "low paying jobs", others require some kind of job evaluation and still others prescribe the factors for evaluation schemes. It is far too early to assess the impact. Both England and Canada have laws which approach comparable worth but these laws lack "teeth." Few cases are heard as these laws apply primarily to broadly similar jobs and most women's jobs are not similar to men's at all.

The first step is the recognition that pay inequity exists. It exists not because of the job but rather because of the gender of the people who traditionally held the jobs. Next, the public sector can be a model for the private sector. Much of the activity in government comes from the vulnerability of public officials to public opinion in cases involving social justice. Furthermore, the high degree of unionization among public sector (over double of that in the private sector) and labor unions have been unanimous in their support of the comparable worth cause. Given the mandate of their constituencies, public personnel managers can be on the vanguard of major changes in pay systems. The forces supporting change appear strong and public human resource managers are faced with the opportunity to provide working models for initiative in the private sector.

REFERENCES

American Society for Personnel Administration, The American Compensation Association Elements of Sound Base Pay Administration (ACA: Scotsdale, Arizona) 1981.

Blumrosen, P. I. "Wage Discrimination, Job Segregation, and Title VII of the Civil Rights Act of 1964" Journal of Law Reform 12 (1979)

Brown, Marsha "Getting and Keeping Women in Non-Traditional Jobs" Public Personnel Management 10 (Winter 1981) 408-411.

Cascio, Wayne Applied Psychology in Personnel Management 2d. ed. (Virginia: Reston Publishing Co.) 1982.

"Civil Rights Commission Chief Calls Pay Equity a 'Looney' Idea" Resource (December, 1984) p. 1.

Collette, Catherine O'Reilly "Ending Sex Discrimination in Wage Setting" Proceedings of the Thirty-Fifth Annual Meeting of the Industrial Relations Relations Research Association (Madison, Wisconsin: IRRA) 1982, 150-155.

"Comp Worth: $60 billion plus Price Tag" Action 17 (June 1981) p. 1.

Davis, Keith Human Behavior At Work (New York: McGraw Hill) 1981.

England, Paula "Women and Occupational Prestige: A Case of Vacuous Sex Equality" in Signs, Journal of Women in Culture and Society 3 (1979).

Flick, Rachael "Undermining The Women's Movement" Human Rights 12 (Fall 1984) 26-29, 51-53.

Green, Ronald M. "Comparable Worth— The Compensation Issue for the 1980s?" Proceedings of the Thirty-Fifth Annual Meeting of the Industrial Relations Relations Research Association (Madison, Wisconsin: IRRA) 1982, 157-161.

Hay Group The Client Briefing, No. 102 (Reward Management Division of Hay Associates) August 6, 1981.

Herzog, Arthur The B.S. Factor: The Theory and Technique of Faking It in America (New York: Simon and Schuster) 1973.

Hurd, Sandra, Murray, Paula and Shaw, Bill "Comparable Worth: A Legal and Ethical Analysis" American Business Law Journal 22 (Fall 1984) 407-427.

"In Minnesota, 'Pay Equity' Passes Test but Foes See Trouble Ahead" The Wall Street Journal May 10, 1985, p. 1, Sec 2.

Johannesson, Russel E., Pierson, David E. and Koziara, Karen S. "Comparable Worth: The Measurement Dilemma Proceedings of the Thirty-Fifth Annual Meeting of the Industrial Relations Relations Research Association (Madison, Wisconsin: IRRA) 1982, 162-168.

Lemons et al. v. The City and County of Denver (17 FEP 906)

Livernash, E. Robert (ed) Comparable Worth: Issues and Alternatives (EEAC, Washington, D.C.) 1980.

Lorber, Lawrence Z., Kirk, J. Robert, Samuels, Stephen L. and Spellman, David J. III Sex and Salary: A Legal and Personnel Analysis of Comparable Worth (Arlington, Virginia: The ASPA Foundation, 1985).

Remick, Helen ed. Comparable Worth and Wage Discrimination Philadelphia: Temple University Press) 1984.

Remick, Helen "The Comparable Worth Controversy" Public Personnel Management 10 (1981) 371-383.

Schwab, Donald "Job Evaluation and Pay Setting: Concepts and Practices' in Comparable Worth and Alternatives Equal Employment Advisory Council, 1980.

"Span of Discretion Measures Job Worth" Public Administration Times (November 1, 1981).

Stencil, Sandra "Equal Pay Fight" Editorial Research Reports 2 (March 20, 1982).

"Supreme Court Rejects Idea of Comp Worth" Resource (December, 1984) p. 1.

Treiman, Donald J. and Hartman, Heidi L. Women, Work and Wages: Equal Pay for Jobs of Equal Value (Washington: National Academy Press) 1981.

U.S. Department of Labor Perspectives on Working Women: A Databook 1980.

U.S. Department of Labor, Women's Bureau The Earnings Gap Between Men and Women 1979.

Wall Street Journal "Letters to the Editor" July 27, 1981.

* * *

Sex and Occupational Differences on the Perceived Importance of Wage and Salary Determinants

Scott L. Fraser and Michael W. Johndor, Indiana University - Purdue University at Indianapolis, Ralph A. Alexander, The University of Akron

The primary focus of this study is the extent to which males and females differ in their perceptions of the importance of various wage and salary determinants. Without such information, it is difficult to predict how fair a given job evaluation method will be seen by those who must live with the resulting pay levels. This information would also be useful in choosing between various job evaluation methods when pay equity is a concern and would serve to focus future studies on the perceived fairness of employment practices.

A secondary focus of this study is the possible existence of occupational differences in the perceived importance of wage and salary determinants. It is not known to what extent various wage determinants are valued by those in different occupations. In practice, organizations commonly use different job evaluation methods or instruments for different job families. Milkovich and Newman (1984) note that employee acceptance is thought to be better when different methods are used for different jobs. It is possible, however, that the use of multiple job evaluation methods may lead to perceptions of inequity. Individuals in low paying jobs may believe that they are being unjustly compensated if their pay is determined using a different system than is used for higher paying jobs. If employees in most occupations were found to agree on what pay should be based, it may be possible to develop one job evaluation system that would be seen as equitable by people in most jobs.

This study will examine occupational differences in the perceived importance of wage and salary determinants in addition to sex differences. Based on methodological considerations noted above, even if occupational differences are found to be minimal, it does not necessarily suggest that organizations should use one job evaluation method for all jobs. Future research would have to establish the practicality of such an evaluation system.

Method

Subjects

Questionnaires were administered to 428 subjects recruited from a variety of settings: managerial, clerical, and blue collar employees of an automotive component manufacturing plant; administrative and clerical personnel from a public school system; employees of a newspaper; and students enrolled in graduate, undergraduate, and continuing education courses at a large, urban university. A total of 370 usable questionnaires were returned.

The wage and salary determinant questionnaire asked the subjects to rate each item twice. One set of ratings (the "Should Affect" ratings) was obtained for how important subjects thought the items should be in determining the wage and salary level for jobs. A seven-point scale, with anchors ranging from "Very Important" (a rating of 7) to "Very Unimportant" (a rating of 1) was used. For the second set of ratings (the "Does Affect" ratings), subjects were asked to rate how important they thought the items actually were in determining the wage and salary levels in most organizations. The same seven-point scale described above was used. Subjects then provided the following demographic information: age, sex, educational level, occupation, and number of years in present job.

Procedure

Subjects were given the questionnaire in groups of 5 to 43. The subjects were told that the study was concerned with their perceptions of the importance of wage and salary determinants. Subjects were also instructed to rate the items based on their perceptions of how the items should affect or do affect the wage and salary level in jobs in general, not for any one specific type of job or for any one organization. Subjects were told that their responses would be anonymous and that the results were intended for research only. Subjects typically required from 15 to 20 minutes to complete the questionnaire.

Analyses were performed to determine whether or not sex differences in the Should Affect and Does Affect ratings of the factors occurred. The results for the wage and salary determinant factors, as well as the results for the individual items, indicate that few ratings differences due to sex occurred. In no instance was there a sex difference of .46 or greater on a 7-point scale.

Analyses were then performed to determine whether or not occupational differences in the ratings existed. For both the Should Affect and the Does Affect ratings of each factor, a One-Way Analysis of Variance (ANOVA) was performed with occupational group as the independent variable.

Only one ANOVA yielded a significant effect for occupational group: the Does Affect ratings for Effort. The mean rating was highest for subjects in Service occupations and lowest for Unemployed subjects. When a post hoc comparison (Scheffe's) was performed on group means, however, there were no two groups significantly different at the .05 level. Clearly, subjects in different occupations did not substantially differ in their ratings of the factors.

Given that both the Should Affect and the Does Affect ratings were not substantially affected by either sex or occupation, the possibility still existed that significant differences might be found between the two types of ratings for each factor or item. Specifically, there may be large discrepancies between the extent to which people think

the items should affect pay and the extent to which they believe that the items actually do affect pay. A comparison of the two sets of ratings was facilitated by the use of identically worded items and similar 7-point scales for both types of ratings. The two sets of ratings only differed in the specific instructions given to subjects. The ratings differed for 32 of the 40 items at the .01 level of significance. Subjects apparently believed that some items (e.g. Potential health hazards) were undervalued, or less important in actually determining pay levels than the subjects thought they should be. Other items (e.g. whether or not the job was unionized) were seen as being overvalued, or more important in actually determining pay than subjects thought they should be.

Inspecting the direction of the mean differences for the individual items yields an interesting pattern of results. Evidently, most (24 of 30) of the content items were seem to be significantly undervalued, while most (7 of 10) of the non-content items were seen to be significantly overvalued.

Overall, the results presented above indicate that the ratings of both the extent to which the item should affect and the extent to which the items actually do affect pay levels did not differ with respect to sex or occupational group. However, the results strongly suggest that many items differ significantly in the extent to which people believe that they should affect pay versus the extent to which they believe that the items actually do affect pay.

Discussion

The results presented above represent one attempt to determine what factors people in our society think pay should be based on, as well as what people think pay actually is based on. While future research needs to be done to replicate the results presented above and to extend them to other occupational groups, the results do suggest agreement in the perceived importance of wage and salary determinants.

With respect to pay equity and gender, it may be possible to design or identify job evaluation systems that would be seen as fair by both males and females. Furthermore, it may be possible to identify situations where various job content or non-content factors that may affect perceptions of pay equity need to be considered before a wage and salary system is installed.

There are two potential limitations to the generalizability of the results of this study that must be discussed. First, subjects were asked to rate the characteristics with respect to jobs in general, not with respect to any one specific job. Subjects may have different beliefs for jobs in general than for specific types of jobs, or for their present job. The authors are currently investigating this issue. A second limitation to the generalizability of the results concerns the nature of the importance ratings. Due to the format used ("Very Important" to "Very Unimportant"), it is not possible to tell exactly <u>how</u> the factors should affect pay.

To summarize, additional studies are needed on the factors that are believed to determine, as well as the factors that actually do determine, pay levels. It is important to recognize that the adequacy of job evaluation procedures can be fruitfully addressed from the standpoint of perceived equity as well as from the standpoint of psychometric adequacy and practicality. The present results suggest that males and females, as well as people in different occupational groups, may have similar perceptions concerning what pay should be based upon.

References

Blumrosen, R. G. (1980). Wage discrimination, job segregation, and women workers. Employee Relations Law Journal, 6, 77-136.

Doverspike, D. D., & Barrett, G. V. (1984). An internal bias analysis of a job evaluation instrument, Journal of Applied Psychology, 69, 648-662.

Doverspike, D. D., Carlisi, A. M., Barrett, G. V., & Alexander, R. A. Generalizability analysis of a point method job evaluation instrument. Journal of Applied Psychology, 68, 476-483.

Fraser, S. L., Cronshaw, S. F., & Alexander, R. A. (1984). Generalizability analysis of a point-method job evaluation instrument: A field study. Journal of Applied Psychology, 69, 643-647.

Lanham, E. (1955). Job evaluation. New York: McGraw-Hill.

Madigan, R. M. (1985). Comparable worth judgments: A measurement properties analysis. Journal of Applied Psychology, 70, 137-147.

Milkovich, G. T., & Newman, J. M. (1985). Compensation. Plano, TX: Business Publications Incorporated.

Otis, J. L., & Leukart, R. H. (1955). Job evaluation: A basis for sound wage and salary administration. Englewood Cliffs, NJ: Prentice-Hall.

Rees, A. (1978). The economics of work and pay. New York: Harper and Row.

Remick, H. (1981). The comparable worth controversy. Public Personnel Management Journal, 10, 371-383.

Treiman, D. J. (1979). Job evaluation: An analytic review. Washington, DC: National Academy Press.

Treiman, D. J., & Hartmann, H. I. (1981) Women, work, and wages: Equal pay for jobs of equal value. Washington, DC: National Academy Press.

* * *

PSYCHOMETRIC AND SELECTION ISSUES (Paper Session)

Chair: Christina L. Valadez, State of Washington

Discussant: Bruce W. Davey, State of Connecticut

The Myth of Proportional Representation

Foster Dieckhoff, Personnel Department, Kansas City, MO

Almost all criteria for determining the existence or extent of adverse impact of an employer's hiring procedures are based upon the premise of what may be called "proportional representation". That is, if a given minority population comprises X percent of the labor pool (e.g., Standard Metropolitan Statistical Area), then ideally X percent of that employer's workforce should also belong to that minority population. In fact, the same reasoning is often invoked to test for adverse impact in individual job classifications regardless of size or expected turnover.

On the surface this proportional representation model (of which the 80 percent rule is a derivation) seems to be based upon solid reasoning. However, many selection practitioners realize that even with carefully constructed "valid" selection procedures the actual minority representation often fails short of what the model predicts. The original scapegoat chosen to explain the discrepancy between predicted and actual minority representation was "validity" or the lack of it. Now, some fourteen years after the Griggs decision, a new vocabulary has emerged to explain adverse impact. Included are such terms as socioeconomic disadvantage, language barrier, and differential validity. All of these no doubt contribute to the problem but none explain it. We have, it seems, been looking for ways to make our data fit a model and no one has questioned the validity of the model of proportional representation itself. A tacit assumption of the proportional representation model is that all of an employer's selection decisions are made at one time from the entire minority population. Clearly this is a gross oversimplification of the dynamic processes actually at work. For example, the division of labor and consequent specialization present in today's environment serve to partition already small minority populations into even smaller subpopulations that, because of number alone, are not statistically capable of producing a viable expected value of competitive candidates for a specific job. In this paper it will be demonstrated that the very process of partitioning any population (e.g., white males) into a majority and a minority (based only upon number) tends to induce discrimination against the minority population. That is, even under conditions of optimal validity and equal probability for selection, the minority population will produce a substantially smaller expected value of successful candidates than does the majority population. Consequently, at least part of the problem lies with the model for determining adverse impact and not with the selection processes themselves.

It will also be shown that even statistically insignificant shifts in the distribution of minority test results (from assumed valid procedures) are likely to have adverse effects upon the possibilities for selection from a minority population.

References

Federal Register 43 (August 25, 1978): 38290-38315.

Griggs vs. Duke Power Company, 401 U.S. 424 (1971).

Maslow, Albert P., Staffing the Public Service. Cranbury, N.J.: Albert P. Maslow, 1983.

Roscoe, J.T., Fundamental Research Statistics for the Behavioral Sciences. New York: Holt, Rinehart, and Winston, 1969.

* * *

Further Support for Validity Generalization: A Test Publisher's Meta-analysis

David A. Dye, Psychological Services, Inc., Washington, D.C.

Since the turn of the decade and into the mid 1980's, there has been a renewed interest in testing. The upward swing in the economy has begun to lead to increased hiring and more testing. But, it is the research in two related areas, validity generalization and utility, which have impacted test use and should continue to foster the use of testing for employee selection.

The application of utility formulas to selection programs has shown that substantial gains in productivity can be realized from using procedures with even modest validity (Schmidt, Hunter, McKenzie, and Muldrow, 1979). This fact makes it imperative that employers be provided with accurate estimates of validity. In view of the differences in average validity coefficients found across researchers' studies, it would seem appropriate to base meta-analytic studies on occupational groupings that are relevant to employers.

The present investigation was undertaken by the test publisher to summarize the validity evidence for one of its widely used multiple-aptitude test batteries, the Employee Aptitude Survey (EAS). Specifically, meta-analyses were performed to examine the validity generalizability of the EAS across occupational groups in the prediction of job performance and training success. Until this investigation, no comprehensive validity information had been summarized on the validity of the EAS other than at the job level in a single job setting.

Sample

Five sources of validity information were identified. These included both in-house and external studies, published and unpublished. In all, 81 studies were located. A breakdown of the sources investigated and the number of studies obtained was as follows: unpublished studies described in the test publisher's existing technical report (31), unpublished validation reports written by the test publisher (44), unpublished studies performed by external consultants or by employers (2), and published studies(4).

Procedure

A data coding scheme was developed to identify and capture information from the studies. The number of validity coefficients per study ranged from 1 to 40. For each study, information was collected on the type of job(s), whether job performance or training success was predicted, which test(s) in the EAS were used, the types and estimates of criterion unreliability, the types of validity coefficients, the sample sizes, and the observed validities.

Participants were drawn from a list of 580 organizations and contact persons with potential validity information which was generated from sales invoices and mailing lists.

For the four other sources of validity information, the publisher's staff searched reports and articles, read test reviews (Buros, 1965 and 1978), and performed a computer search of the published literature.

For the 81 obtained studies, the relevant information was coded on data recording forms. From these studies, five occupational groups were identified: professional, managerial/supervisory, technical, sales, clerical, and skilled/semi-skilled. It was believed that these groups were meaningful for validity generalization calculations and for test user purposes (the occupational groups match the categories on which EAS normative data exists).

A total of 429 validity coefficients were recorded. For studies that reported a validity for more than one type of criterion measure (e.g., work sample, job knowledge test, rating), each coefficient was recorded. For studies that reported validities for several dimensions of a single type of criterion measure, only the overall summary coefficient was reported. In one study in which no summary figure was reported, average test validities across the criterion dimensions were recorded with the sample size being a product of the original sample size and the number of dimensions averaged. This resulted in a total of 420 validity coefficients across 81 studies.

Data Analysis

The recorded information was key entered and written to computer tape. The 420 validity coefficients were separated into studies on job performance or training success and sorted into occupational groups.

Frequency counts of their number of validity coefficients for each EAS as a function of occupation group and type of criterion (i.e., job performance or training success) were computed.

Meta-analyses were performed, after the procedures found in Hunter, Schmidt, and Jackson (1982). Due to the variety of criterion measures used within and across studies, many predictor-criterion combinations contained few studies. Therefore, for this analysis, a meta-analysis was conducted only if there were at least five studies available for a predictor-criterion combination within a particular occupation group. (Average weighted validities were calculated for all instances in which there were less than five studies. These can be obtained from the author.) This decision rule was applied both to studies investigating prediction of job performance and training success. This represents a decrease in the number of studies used by Pearlman et al. (1980) in their investigation of clerical occupations in which they required 8 studies for training success and 10 studies for job performance.

From the decision rule, 22 meta-analyses were performed, 12 for job performance and 10 for training success. For all analyses, the criterion measures used were supervisory ratings of job performance and grades or achievement test scores for training success.

For each of the 22 distributions of observed validities, the mean ($\bar{r}$) and variance (S_r^2) of the observed coefficients were computed, with each coefficient weighted by its sample size. In addition, the amount of variance expected by sampling error (S_e^2) was computed. Corrections to the observed variances were then made by subtracting the respective sampling error variances. According to Pearlman et al. (1980), this correction is a conservative one for validities that are not derivations of the Pearson product-moment correlation. Twenty-three percent of the coefficients based on job performance and 23% of the coefficients based on training success were either biserial or tetrachoric r's. The percentage of variance accounted for by sampling error and the residual standard deviation, the square root of the observed variance corrected for sampling error, were also computed.

Corrections to the mean and variance of the residual distributions for criterion unreliability were made, using expected values of the artifact distributions generated by Pearlman et al. (1980), .60 for job performance and .80 for training success. (Consistent with other researchers, only 13% of the studies reported reliability estimates of criterion measures.) The correction to the mean ($\bar{\rho}$) represents the best estimate of the EAS validity after correction for sampling error and criterion unreliability. The variance correction was used to calculate the 90% credibility values. This value represents the 10th percentile in the distribution of true validities, or the point above which 90% of all true validities would be expected to lie. Corrections for range variation were not made as the same predictor (the EAS) was used in all studies and because there was no evidence to

indicate that there was any restriction on the predictor for incumbent samples. The results of the meta-analyses for job performance appear in Table 3; training success results are presented in Table 4.

Results and Discussion

The first set of results provided a historical perspective on the use of the EAS. With the development of the EAS in 1963, clearly the majority of validation studies and subsequent testing has been in nonprofessional occupations. The percentage of collected validity studies to predict job performance is 12% for technical jobs, 27% for clerical jobs, and 46% for skilled/semi-skilled jobs; for studies of training success, the majority of studies are in the technical (69%) and semi-skilled (20%) occupations.

Support for validity generalization of the EAS was found in all instances. The average proportion of variance accounted for by sampling error, weighted by the number of studies, was 80% for the job performance analyses and 55% for the training success analyses. In 10 of the 12 job performance analyses, more than half of the variance in validities was accounted for by sampling error, with 100% of the variance accounted for in five of the analyses. For six of the 10 analyses on training success, sampling error accounted for over half of the observed variance.

When corrections for criterion unreliability are made, all of the 90% credibility values are above zero. This is the important consideration for employers; that is, no local validation study would be necessary for the test-occupation combinations investigated.

In this investigation, a requirement for a minimum of five studies was used for each meta-analysis. Representing a decrease from the number of studies used by Pearlman, et al. (1980), this provides further support for validity generalization using a fewer number of studies. Schmitt, et al. (1984) suggested that the corrections for sampling error using the procedures of Schmidt, Hunter, and Jackson (1982) may be inappropriate when the number of studies is less than six. However, the sampling error correction assumes that studies are conducted independently. Thus, this claim does not seem warranted.

A final discussion centers on the average validities. Although the validities are slightly lower than other reported validities of cognitive ability tests, they reflect the intended use of the EAS and other multi-aptitude batteries. It is a well known fact that selection programs benefit from using homogeneous, uncorrelated predictors. With all of the tests in the EAS being speeded with relatively few items, validity is maximized by forming batteries of appropriate tests (Ruch and Ruch, 1980). At this time, further work is underway to determine generalized battery validities.

References

American Psychological Association, Division of Industrial-Organizational Psychology (1980). Principles for the validation and use of personnel selection procedures: Second edition. Berkeley, California: Author.

Barrett, G.V., Phillips, J.S., & Alexander, R.A. (1981). Concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 66, 1-6.

Bemis, S.E. (1968). Occupational validity of the General Aptitude Test Battery. Journal of Applied Psychology, 52, 240-249.

Buros, O. (Ed.). (1965 and 1978). Mental Measurements Yearbook. (Sixth and Seventh). New Jersey: Gryphon Press.

Hunter, J.E. Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitute Test Battery. Report prepared for the U.S. Employment Service, Department of Labor, 1981.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). Meta-analysis: Cumulating research finding across studies. Beverly Hills: Sage Publications.

Pearlman, D., Schmidt, F.L. & Hunter, J.E. (1980). Validity generalization results for test used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Psychological Services, Inc. (1984). Technical manual: PSI Basic Skills for Business, Industry, and Government. Washington: Author.

Ruch, F.L., & Ruch, W.W. (1980). Employee Aptitude Survey Technical Report. Los Angeles: Psychological Services, Inc.

Schmidt, F.L., Gast-Rosenberg, I., & Hunter, J.E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Schmidt, F.L., Hunter, J.E., & Caplan, J. (1981). Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 66, 261-273.

Schmidt, F.L., Hunter, J.E., McKenzie, R., & Hirsch, H. Questions and answers about validity generalization and meta-analysis. Unpublished manuscript, 1983.

Schmidt, F.L., Hunter, J.E., McKenzie, E., & Muldrow (1979). The impact of valid selection procedures on workforce productivity, Journal of Applied Psychology, 64, 609-626.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta-analyses validity studies published between 1964 and 1982 and the investigation of study characteristics, Journal of Applied Psychology, 37, 307-422.

* * *

An International Perspective of Personnel Selection Systems: British vs. American

Priscilla J. Hambrick-Dixon, New York City Department of Personnel, Bureau of Examinations

Overview

In recent years, New York City's long history of and experience with Civil Service employment testing of large multi-ethnic and multi-racial candidate populations has generated much international attention. One possible reason for this is that several relatively racially and ethnically homogeneous European countries are currently experiencing an influx of immigrants -- from American, Asian, Caribbean and Middle Eastern countries -- seeking employment. Thus, for the first time, Europeans must face the challenges of selecting a multi-ethnic and multi-racial workforce.

The purpose of this paper is to compare the personnel selection systems of Great Britain and the United States to determine whether there are tenets of employment selection which may be essential to assure equal opportunity for and fairness in selection of a multi-ethnic and multi-racial workforce.

Invitation to Great Britain

In early November, Dr. Judith Piesco, Deputy Personnel Director for Examinations and I were invited by the British Commission on Racial Equality (CRE) -- to participate in conferences on employment selection in London, England and Cardiff, Wales. These conferences were attended primarily by personnel officers from the public sector and training specialists from both public and private sectors.

The general purpose of our visit was to share how New York City has attempted to meet the challenges of employment selection with its large and diverse ethnic and racial candidate population. As requested by the British Commission on racial equality, the specific goals and objectives of our visit were to: 1) discuss the American legal mandates and professional standards for the development, validation and administration of employment selection procedures; 2) explore how the U.S. deals with discriminatory practices and particularly those on the basis of race; 3) discuss the various modes of employment assessment used in the United States and the research on validity and efficacy thereof; and, 4) reflect upon how racism impacts upon the white majority and other ethnic minorities (blacks in particular) in the employment arena.

Since this was our first visit to Britain, we were concerned that we were at a disadvantage and that our observations and perceptions of the nature and scope of the personnel selection system in Britain in comparison to New York City's system might tempt us to overgeneralize. To guard against this temptation, the following questions were focused upon: 1) From what framework can one view and evaluate the

American and British personnel selection systems? 2) Within what socio-political-context should the two personnel systems be viewed and evaluated as well? 3) On what specific dimensions should these two personnel systems be compared? 4) How are discriminatory practices in employment defined, perceived and dealt with in Britain and the United States (New York in particular)? It was evident that this visit would provide an invaluable opportunity for gaining an international perspective of employment selection and evaluating the strengths and weaknesses of the American employment selection system.

Comparative Analysis of British and American Personnel Testing Systems

A comparative analysis was made of the employment selection systems of Great Britain and the United States (New York City) -- on thirty-one dimensions -- to ascertain how the two systems view employment issues related to selection of multi-ethnic and multi-racial job candidates. Parenthetically, these dimensions have been viewed on a continuum and from a developmental perspective based upon American civil rights events and developments. The thirty dimensions were categorized into four major areas: psychometric, legal, social-political, and economic, as indicated in Table 1.

The psychometric dimensions included: modes of assessment; dimensions to be measured; objectivity vs. subjectivity; value of particular modes of assessment; emphasis on performance; type of measurement; unit of measure; basis for inference about job performance; derivation of criteria or standards for performance; evaluation of outcomes; and relationship to professional organizations concerned with testing issues. In comparison to the United States -- on these dimensions -- in Britain: 1) most employment selection procedures are more unimodal and unidimensional; that is, the selection interview -- the least reliable and valid -- emphasizes the assessment of oral communication skills; 2) the types and units of measurement are generally more qualitative, subjective and impressionistic relying upon inferences about covert rather than overt performance; 3) the criteria, standards for and evaluations of performance also evolve more from the subjective realm; 4) little relationship exists between the professional educational and psychological associations and the legal professions.

The legal dimensions encompassed: laws governing employment selection practices; mandates concerning the job relatedness of selection procedures; legal sophistication of job candidates; burden of proof regarding allegations of racial discrimination; the culprit and responsibility for racial discrimination in employment; mandates concerning the level of skills required at job entry-level; and mandates for recruitment. On the legal dimensions, compared to the United States, in Britain: 1) a Code of Practice and Race Relations Act (1976) exists concerning employment practices, however, these are not enforceable legislations; thus, employment selection procedures may lack job-relatedness; 2) requirements tend to be set at the highest level of entry-level skills rather than minimal level; 3) no recruitment of or affirmative action on behalf of minorities is required; 4) the burden of proof lies with

the job candidate when he/she believes that he/she has been discriminated against; 5) the individual job candidate is perceived to be the culprit and is responsible for being discriminated against and 6) job candidates are less litigious (Carby and Thakur, 1977).

The socio-political dimensions included: tolerance for diversity; value of multi-ethnicity; aggregation vs. individualism; socio-political context; commitment to the establishment of a multi-ethnic workforce; perceptions of the bases of racism; terms used to describe particular races; impact of racism; size of the population and country; emphasis on racial awareness and training. In comparison to the United States, in Britain, there seems to be: 1) a much stronger preference for homogeneity of races and ethnic groups; 2) little momentum of a civil rights movement; 3) a stronger tendency for aggregation; 4) less expressed commitment to the establishment of a multi-ethnic/multi-racial workforce; 5) more emphasis on superiority of whites and inferiority of black as basis for racism as indicated in newspaper articles and letters to CRE; 6) a reference to groups with dark skin color as "coloreds"; 7) a smaller country and population yet higher demand for jobs; 8) a stronger emphasis upon racial awareness and training; 9) less use and impact of the media for publicizing racial discrimination.

With regard to the economic dimension, the British economic system is more socialistic, while the United States if more capitalistic.

According to the British Commission on Racial Equality, in Britain, the employment of all protected classes is problematic. But for the Black race, it is considered very serious (Carby and Thakur, 1977). The observations described in the analysis above have sparked the burning question: To what extent is the status of equal opportunity in employment in the United States today, influenced by social equality ideologies, EEOC Uniform Guidelines (1978, 1979), professional standards and collaboration between our legal and professional organizations?

Implications and Conclusions

Considerable attention has been directed recently, toward federal regulations to prevent unfair discrimination in employment testing (APA MONITOR, 1985). The Equal Employment Opportunity Commission is currently reviewing the Uniform Guidelines in light of allegations that they are technically outdated and problematic for employers. Moreover, issues related to the definition and legality of affirmative action (as it relates to quotas) has become a major area of controversy.

The British are very critical of "positive discrimination" (the practice of giving preferential treatment to racial minorities purely on the grounds of race-quotas) (Holland and Parkins, 1984). Many American intellectuals, unions and government officials are equally concerned about the current false assumptions and conceptions of equality which may be contradictory to the ideals of equality of opportunity for all regardless of race, creed, color, religion, handicap, nationality or national origin. Many scholars have argued that the American experience with positive or reverse discrimination is an example that Great Britain should not follow (Holland and Parkins, 1984).

Overall, this analysis reveals that the American ideological basis for equality in employment evolved from good intentions but may be somewhat naive from a world view and in the context of practicality. The logistics of promoting such ideals will continue to be a matter of inquiry and controversy for a very long time. Perhaps, there may be lessons to learn from the British as they meet the challenges of selecting a multi-ethnic and -racial work force.

References

Corby, K. and Manab, Thakur. No Problems Here?. London: Institute of Personnel Management, 1977.

Cordes, Collen, Review may relax job testing rules, APA Monitor, Vol. 16, No.5, 1985.

Equal Employment Opportunities Commission, Civil Service Commission, Department of Labor and Department of Justice. Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. Federal Register, 1978, 43, 38290-38315.

Holland, Kenneth, M. and Parkins, G. Reversing Racism. Social Affairs Unit: London, 1985, Research Report No.5.

Sowell, T., "Weber and Bakke, And the Presumptions of Affirmative Action" in Discrimination, Affirmative Action and Equal Opportunity. Vancouver, The Fraser Institute: 1981.

* * *

THE USE AND MISUSE OF ITEM BIAS STATISTICS (Symposium)

Chair: John G. Veres, Auburn University at Montgomery

Discussant: Keith Pyburn; McCalla, Thompson, Pyburn and Ridley

Difficulties with Delta

John G. Veres, III and Mary Anne Lahey, Auburn University at Montgomery, Alabama

There seems to be considerable consensus of opinion in the professional literature that there might be some difficulties with the method for detecting item bias. If one looks at Title VII litigation, Angoff's Delta seems to be a technique that experts frequently rely upon. This method is also sometimes called the Transformed Item Difficulty Method (TID).

Angoff's Delta is accomplished in 5 steps. The first step is to calculate item difficulties as one ordinarily would compute them in a standard item statistics procedure. The next step is to take these item difficulties

and transform them into normal deviates. If one selects a mean of 13 and a standard deviation of 4, the deviates are referred to as deltas. One would then take the deltas for 2 groups of interest, in our case we were looking at the blacks and whites, and plot for each item a pair of deltas, placing one on the ordinate and the other on the abscissa. When finished, an elliptical plot results.

The major axis of this ellipse is an indication of the average difference in difficulty level over the course of the test for the 2 groups. Therefore, items that fall at some distance from the major axis are relatively more difficult for one group or the other. There are a couple of different methods for looking at distance from this major axis, some of which drop parallel with the ordinate and some which drop perpendicular in the major axis. The latter method is the most widely used. Angoff states that these types of items have a different psychological meaning for one group than the other.

The advantages of the Delta plot may be summarized as follows. First, it is a very simple procedure and is relatively easily explained. It has been stated that the Delta does not require much data to achieve stability. It is also inexpensive. Lastly, it is something anybody could do in 2 or 3 minutes using a canned statistical package.

There have been some problems identified with Delta, however. One of those is that there is some confounding of the Delta value item discrimination. That is, items which do a good job of discriminating individuals in the sample are likely to be identified as biased by this procedure. The removal of the items that would be identified as biased would, therefore, reduce the mean point biserial correlation for the test. This is not desirable. A corollary of this would seem to indicate that items with middle to high difficulties would be removed since they tend to be most discriminating.

Our study focused on the performance of Delta over several administrations of a test where we had a relatively large number of people. The test that I will be talking about is a 120-item multiple choice test which has been administered a fair number of times and we have data today on 9 of the administrations. The racial composition for the first 9 administrations of the test was 516 blacks and 4,841 whites. Thus, we have a substantial number of both blacks and whites on which to do Delta plots. There are a number of things about this test which will help in evaluating our results. One of the things is that this test had substantial adverse impact, so, if Delta is indeed a good index of item bias, one would think that there would be a substantial number of items which Delta would identify as biased.

The first thing that we examined was the cross tabulation of Delta with item difficulty. We looked at both the p-values generated by the black sample and the p-values generated by the white sample. This analysis indicated that the exam has substantially more difficult items for blacks than the sample of whites. Contrary to the expected pattern, items were identified by Delta as biased, did not tend to be more difficult.

The removal of the "biased" would not significantly affect the relative difficulty of the test. Even though there are many more hard items for blacks, Delta is not identifying them. In fact, there are fewer items that Delta identified as biased against blacks which were relatively more difficult for blacks than those for whites. Removing those biased items would not dramatically change the difficulty of the examination.

Given these results, when we considered cross-tabulation of Delta with item discrimination, we found that there is not a substantial amount of difference. If anything, removal of the biased items would have slightly aided discrimination so, given these findings, the classic literature-based criticisms of Delta don't seem to be born out. In fact, one could have almost generated these numbers with a random table.

Having not found much interest in the cross-tabulations, we decided to track an items Delta value across multiple test administrations. If we had an item that was used on at least 4 administrations, we included it in our analysis. The way Delta is plotted assures that on any given test, half the values will be positive and half will be negative, so we wanted to track the stability of these signs across time. Over the nine test administrations the values varied considerably from one administration to the next. We looked at the 5% significant level and we found that the Delta identified relatively few items, 38 out of 137, as biased, slightly favoring whites over blacks. We next relaxed our alpha level to 10% and we found that our results were due to the small sample. When we backed off the Type I error rate so that 6 to 3 splits on 9 administrations were identified as biased we still found an essentially random pattern. Tracking across multiple administrations, items favored blacks just about as many times as they favored whites. For all intents and purposes, at least in this particular sample of data, Delta was so unreliable that we just couldn't get anything meaningful from it at all.

What do our findings imply for test construction? Given the unreliability of this particular index, it seems to us that it is <u>not</u> good practice to compute Angoff's Delta on a given test and eliminate test items which appear to be biased against blacks or whichever group is in question. The procedure will probably eliminate the items which may well favor that group on subsequent administrations.

* * *

Practical and Theoretical Applications of Item Bias Studies

Chester I. Palmer and Wiley R. Boyles, Auburn University at Montgomery, Alabama

In the particular study we did, as a result of an agreement between a state personnel department and the U.S. Department of Justice, the personnel department agreed to conduct a criterion-related validity study of a test which had been used to select entry-level employees in the job classes Clerk, Clerk Typist, and Clerk Stenographer. The test was a 100 item multiple-choice test built on the basis of a job analysis which indicated that the principal tasks were filing and retrieving, proofreading, and a general communications task primarily involving taking and relying messages; the test was an earlier version of the test which Ron Downey discussed. The test did not cover specialized typing and stenographic skills, which were assessed in another part of the selection process.

We used several different criterion measures in the study, but two kinds of measures predominated. One set of measures consisted of supervisory ratings, some obtained from the standard supervisory evaluations used for employee evaluation, and others obtained under research conditions by members of the study team after training the supervisors in employee ratings. I am not going to talk very much about those today, because from many points of view, the more interesting set of criterion measures consisted of scores on a performance test consisting of simulations of job tasks. The performance test involved three ten-minute filing tasks, one requiring unspeeded filing, one highly speeded filing, and one speeded information retrieval. The performance test also involved two kinds of proofreading tasks, one finding errors in letters and memos and one checking for transcription errors between handwritten copies of forms and typed versions. Thus we had an unusual advantage in that the results of these simulations gave us a set of measures which we could use to estimate the likely effect on validity of various changes in the selection test. We were fortunate enough to have large numbers of subjects, although we had a low selection ratio: over five thousand applicants took the selection test, approximately seven percent of whom were hired. We have our job simulations to 184 people hired on the basis of the test.

In addition to item-level analyses, we also performed analyses of the test as separated into five subtests based on the KSAs which the items were intended to measure: English language, alphabetizing, following instructions, communications skills, and all other topics. We were surprised to find that racial differences were largest on the subtest on following instructions, and second largest on the subtest on other topics. Differences were smallest on alphabetizing.

Using a random sample of papers from the original test, we applied several different item analysis procedures based on relative performance by black and white applicants in an attempt to select approximately 70 items from the original 100 to form a new test which would have validity

comparable to that of the old test, but lower adverse impact. We constructed three such new tests. For cross validation, we then took a second random sample of the original papers, disjoint from the first sample, and compared the characteristics of the new tests.

In general, the results were mixed. The shortened version of the test was actually a better predictor of filing performance than the original test, probably partly because since racial differences were smallest on alphabetizing, such items were generally retained on the shortened test, where they formed a greater proportion of the whole. The shortened version was not as good a predictor as the original test for one of the parts of the proofreading performance test. On balance, the tests were of comparable validity. The correlation of scores on the new test with race was substantially lower than that for the original test.

For realistic selection ratios of 10-20%, the new test did not eliminate adverse impact, but it improved the adverse impact ratio by about .15. We believe this is a worthwhile improvement. In addition, there is reason to hope that further applications of such methods might reduce adverse impact even further, although inspection of response data indicates that it is unlikely that any such method will completely eliminate the problem. Nevertheless, one clear result from our study is encouraging: at least in this case, it was possible to use item-bias analyses to lessen adverse impact while maintaining validity.

We now turn to more general questions. We believe that it is fair to say that the single biggest problem with the results of item-bias studies is the difficulty of interpretation. Our colleagues have already pointed out that those analyses which might seem the most penetrating, such as those using item characteristic curve methods, are often impractical because they require unrealistically large numbers of subjects and extensive human and computer resources. Simpler methods such as Angoff's delta and the chi-square analyses can be applied more often, although the results are often unstable in small samples. But the basic question still remains: If any method suggests that a particular question is biased, what should we do about it, especially realizing that in content-oriented test construction we usually have no way to determine the effect on test validity of removing such questions?

We realize that there is some controversy regarding this issue, but we take the following position: If we depend on content-oriented test construction, then ultimately it must be the content of a question that determines its suitability for use on a test. We do not believe any item should be discarded <u>solely</u> on the basis of item-bias statistics. On the other hand, bad item-bias statistics should be a warning that we should examine the question closely, in the same sense that bad reliability statistics are a warning. In addition to looking for obvious biasing factors, the first step is to compare the suspect item with other items intended to measure the same KSA, with special regard to the form of the items. If the bias statistics are consistently worse for one type of item than another (say worse for items asking which word is spelled correctly than for those asking which is spelled

incorrectly), we should clearly use the form with the better bias statistics unless there is some compelling reason to believe that the other form has higher validity, such as very high point-biserials for both groups separately and for the combined group. (Note that some item-bias methods do tend to assign high bias to extremely discriminating items.)

If there are large differences on all types of items regarding a particular topic, we can try to think of new kinds of items. Especially when the test is not homogeneous, however, it is important to remember the possibility of real relative differences. For example, an item analysis over the entire ACT battery would show many mathematics questions with bad bias statistics; traditionally, racial differences are largest on the Mathematics subtest and smallest on the English subtest. A major part of the difference is caused by the fact that black students traditionally choose fewer mathematics courses than do whites; but nearly all students are required to take four years of English. In this case, the item-bias results are detecting a real difference in relative performance. Under present circumstances, it seems likely than any reasonable mathematics test would have many items with bad bias statistics when analyzed with the full battery. In this kind of situation, it would be more meaningful to analyze only the mathematics subtest. Often a two-level analysis is the most meaningful: First, consider racial differences in performance between the clusters of items intended to measure particular KSAs; then do the bias analysis on the clusters. It is not always necessary to do any rescoring. Using Angoff's delta, for example, relatively uniform delta values within a cluster of items suggest the possibility of real differences. This kind of analysis at least gives some information regarding which problems may be caused by real differences and which may be artificially produced by particular items.

When we applied this kind of analysis to our clerical data, we became especially interested in the cluster of items on alphabetizing. Although the cluster as a whole had less adverse impact than the other clusters, we were surprised by the amount of variation within the cluster. Some of the variation had obvious causes -- items dealing with names beginning Mc or involving non-English prefixes. But on many other items, the cause was not obvious. We believe that this situation illustrates one of the major problems with interpreting such item-bias results: We have no theoretical context. If we knew what it was about an item that caused it to show high or low bias, we could then make a rational judgement whether to replace the item or to retain it because we believe there is a job-related difference in performance. We spent some time trying to constuct such a theory in the case of these alphabetizing items. We eventually devised the theory that black applicants were relying on sound to a greater extent than white applicants, who seemed to rely more on sight and by sound, two were difficult by sight but involved common names which were either very difficult to pronounce or for which the pronunciation did not agree with the alphabetization.

We have now studied the results of the new test containing those items, and they do not particularly confirm the theory, although they do not strongly contradict it either. Unfortunately, it is very hard to test

a theory using only five examples. But whether or not the theory is correct, there was one interesting result: On the new test, for both blacks and whites, the experimental item with median difficulty involved unpronounceable names, even though each name began with a different letter. Despite our apparent lack of success, we have learned from the attempt and we intend to keep trying. We would urge you, when confronted with differences in item-bias between apparently similar items, to formulate and test possible interpretations. Once we understand the source of such differences, we will be in much better position to construct tests which are both valid and fair.

In more general terms, we believe that item bias studies can be helpful in test construction and validation. There are legitimate concerns involving sample sizes, unreliability, and other problems with the methods. But we can profit from such studies if we remember that using the results is a matter of professional judgement, rather than devising procedures that automatically discard items based on their bias statistics.

* * *

Application of a Latent Trait Approach to Detecting Item Bias

Ronald G. Downey, Kansas State University

Definitions of test bias have proliferated at almost an exponential rate since the 60s and while some single approach has emerged to eclipse all the others. Most approaches to test bias have dealt with bias as an issue of validity differences between subgroups rather than using the egalitarian hypothesis that suggests that mean differences between subgroups are a prima facie case for bias (Shepard, 1982). Others (e.g., Hunter, Schmidt, & Hunter, 1979) have suggested that most, if not all, differences between validity coefficients are due to statistical problems associated with small samples and large measurement errors.

While at one level the major approaches to detecting item (as distinct from test) bias differ from the methods used in test bias research, they are not inconsistent with a focus upon the construct validity of a test (Shepard, 1983). Shepard (1982) in her chapter on definitions of bias, identified two major classes of item bias methods, logical and empirical.

Logical methods are primarily dependent upon the use of judges to determine the "bias" in an item. These methods are very consistent with the content validity approach with the emphasis upon developing items which have the content and structure inherent in the definition of the test. Tittle (1982), in her discussion of judgemental approaches, identifies a variety of concerns that judges can be asked to consider including sterotyping, etc.). Another area judged is the degree to which items will be expected to lead to differential group performance.

Burrill (1982) concludes, however, that judgements of items that will show groups differences do not identify the same items as identified by empirical methods. Thus, the use of judgemental approaches to item bias would appear to be invaluable during the writing and development of items but have limited utility for identifying items which are differentially responded to by subgroups.

The empirical methods cover a broad range of techniques, methods, and theories. Almost all of the data based methods depend upon an internal criterion(a) and, therefore, if the test as a whole is biased, then the item biased methods generally state that they may not be able to detect item bias. The circular nature of empirical methods is a weakness that can only be dealt with in a limited fashion. Perhaps the most theoretically sound method to studying item bias is item response theory (latent trait theory). Given the complex nature of item response theory it would be appropriate to spend some time discussing the major components of the theory.

Traditional approaches to testing have hypothesized that the relationship between the response to an item and the underlying (latent) trait should be positive, if the item is a part of the trait. The positive relationship had been generally assumed to be linear, and therefore, dependent upon correlational indices between item responses and the total test score. The use of these techniques are applied to multiple choice items that are scored correct/incorrect and yield acceptable results (see chapter 15 in Lord and Novick, 1968). In addition to a concern with the nature of the relationship between the items and the underlying trait, is a concern with two other parameters, item difficulty and guessing levels. Item difficulty is an estimate of the average response to the item by the sample. The guessing factor is a concern with individual's opportunity for getting an item "correct" with little or no knowledge of the item. Figure 1 shows the linear model (Torgerson, 1958), generally assumed in traditional measurement theory. The model has several limitations that distract from its utility. Further, as Lord and Novick (1968) have pointed out, the estimates of the important item parameters are sample dependent and can change rather drastically from sample to sample. While a variety of other item characteristic curve models have been suggested, the most useful models have assumed an S-shaped type of function. Of this family of curves the logistic function (Birnabaum, 1968, p.399) is both convenient and has been the most researched and accepted model. The logistic curve displays the relationship between the trait and the probability of a correct (positive) response to an item.

The intent of this paper was to apply procedures used in logistic theory to a set of test items and check for bias.

Method

Subjects

Subjects were a sample of 2,675 individuals taking a state employment examination for clerk typist/stenographer positions. There were 2,639 females and 35 males (one individual did not mark his/her sex).

Less than one percent indicated they had not finished high school, 22 percent indicated they had graduated from high school, 33 percent indicated they had had business training, and 44 percent indicated some college or a college degree. After exclusion of individuals who indicated a race other than white or black and individuals who did not complete all items in the examination, 797 blacks and 1864 whites were left in the sample.

Materials

The examination was composed of 100 items. It was built to measure four tasks as identified by job analysis: knowledge of the English language (grammar, composition, sentence structure, and spelling); knowledge of the English alphabet (filing); ability to follow written instructions; and ability to interpret factual information. While all items were multiple choice, the number of options varied from 2 to 5. The content of the items was varied and included items on filing, spelling, use of the dictionary, reading comprehension, practical job knowledge, etc. There was no time limit on the examination.

Procedures

Subjects' responses to each item were scored as correct, incorrect, or blank. The thetas and item parameters were estimated using the LOGIST program provided by ETS (Wingersky, Barton, & Lord, 1982).

Results

The reliability analysis of the entire data set yielded a coefficient alpha of .815, F (2548,252252) =5.41, p less than .0001. The overall mean was 78.24 and standard deviation was 8.32. The average inter-item correlation was very low, .046. Nine items were found to increase the reliability of the test when they were removed. The items with decreasing reliability were, in most cases, not the very easy items.

The mean for blacks was 74.13 (s.d.=8.78) and the white's mean was 79.57 (s.d.=7.84). This difference was significant, t(2659) = 44.07, p less than .0001. The coefficient alpha was .811 for blacks and .803 for whites.

The test appeared to be moderately easy for the majority of individuals and there was a substantial difference between racial groups. On the surface the test is a prime candidate for further work to investigate item bias.

The total sample of individuals was analyzed using the LOGIST program. All three item parameters and theta were estimated and processing was stopped after the first two steps of the program with standardization on the theta values. The resultant c values were fixed for both groups and the blacks and whites were run separately with the thetas a and b parameters being reestimated through the full 4 steps. The final a and b values for each group were then tested for differences using the procedures outlined by Lord (1980).

The c values estimated from the combined run were, with the exception of 11 items, set by the program at a value of .2134. Of the eleven other items, 10 c values were found to be greater than .2134 and 1 was less. For the black group, the a values tended to be low, only 4 were greater than 1.0 and the b values were predominantly (76%) negative. A similar picture emerged for the whites, with only a values exceeding 1.0 and 83% of the b values being negative. These values are indicative of a test with a preponderance of items that are easy and differentiate primarily at the lower end of the score distribution. The analysis of the test for the blacks yielded two items where there were difficulties estimating the b parameters and for whites there were four items where the b parameters were difficult to estimate. Table 1 gives the a, b and c parameter estimates for both groups for all items.

Of the 100 chi squares calculated, 69 were significant at the p less than .05 level or less and 31 chi square values did not reach significance. The variances used were estimated from the 2 X 2 information matrix.

Discussion

The raw score difference between the two groups was quite large, given the limited variance of the test and there was some evidence that variability in the scores for the blacks versus the whites differed. The item response curve analysis yielded a preponderance of items which were biased.

The real problem would seem to rest in the degree to which the original test was unidimensional. The ICC approach assumes that the test is in fact unidimensional and if this property is violated to any great degree then the procedures and findings must be questioned. There is good reason to suspect that the original set of items was unidimensional. The combination of reading comprehension items with spelling items with practical job knowledge items would lead one to think that several types of ability were being measured. The differences between blacks and whites were not distributed equally over the different types of items. For the 5 dictionary usage items, all showed significant differences between blacks and whites. For the 18 reading comprehension items, 16 were found to be different. For the 28 filing items, 19 were different. For the 34 English usage items, 22 were found to be different. Finally, for the 15 practical job knowledge items, 6 were found to be different. When a chi square was computed for the type of question by significance findings, it yielded a marginal finding, chi square(4) = 9.38, p less than .10.

Given the large number of chi squares computed it is easy to forget what is being tested and the meaning of the tests. As a general rule the black sample's b parameters were higher (less negative or more positive) than the white sample's estimates. The results for the a parameter yielded more variability in the direction of the differences. The items on the whole were shown to be moderatly effective in differentiating

between various ability levels (primarily at the lower end of the scale). The picture that emerges from these tests is not dramatically different than that shown by the initial item difficulties and item discrimination values. The test is most effective in differentiating between individuals at the lower end of the ability distribution and, given the mean differences between the subgroups, does this better for the blacks.

The major difficulty in trying to understand the results of this study lies with the degree to which employment tests meet the necessary assumptions of item response theory. The assumption of unidimensionality should be seriously questioned in this case. Certainly the item content was drawn from several different domains and the average correlation among the items was .046. Inspection of the correlation matrix did not reveal any large negative correlations nor any large positive correlations. The low average correlation represents a set of items with generally low relationships. The items were restrictive in variance and this was due to many of the items being very easy. This restriction was a potential factor in the low intercorrelations. The evidence of subfactors within the test would be difficult to obtain with this type of correlation matrix.

The low difficulty level of the items also leads to another problem: the estimates of the b parameters were often so far down on the theta scale as to make them very unreliable. This was reflected in the estimates of the variances for the b. If the estimates of the parameters were not good, then the assumptions of the chi square test may not be met.

Given the failure of the test to meet the required assumptions of item response theory, the question of biased items can not be clearly answered in this study. The results do, however, raise several questions about the test and its characteristics and suggests that the developers/users should be concerned about the potential for bias occurring. More specifically, before considering the use of the Lord (1980) chi square approach, researchers/users should attend to the following concerns:

1). Programs to estimate item response parameters are both costly and difficult to use. This study required approximately $4,000 worth of computer time to put the program on the system and to run the data.

2). The estimation procedures require both large numbers of subjects and items.

3). The requirement of unidimensionality is a serious problem for many employment tests.

4). The chi square test rises and falls on the estimation of the item parameters. The estimation procedures are difficult and sometimes are not very accurate.

5). Given the complex nature of the calculations and the abstract nature of the information used in the bias analysis, it may be very difficult to use the chi square type of analysis in dealing with either the public or the courts.

6). The procedures still lack adequate exploration.

7). Finally, as with many multivariate procedures, the chi square test tells us there is a difference but does not help us to understand where (in the a or b parameter).

While the item response theory approach to the study of item bias is most likely the soundest method available, its utility is based upon our ability to meet its rather stringent requirements and assumptions. At the present, it is very questionable if these requirements can be met in most employment testing situations. It is likely that the future will see improvements and advances in this field that may make its techniques more available and useful to industrial applications.

References

Berk, R.A. (Ed.). (1982). Handbood of methods for detecting test bias. Baltimore: The John Hopkins Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick (Eds), Statistical theories of mental test scores (pp. 397-549). Reading, MA: Addison-Wesley Publishing Co.

Burrill, L.E. Comparative studies of item bias methods (1982). In R.A. Beck (Ed), Handbook of methods for detecting test bias. (pp. 161-179). Baltimore: John Hopkins U. Press.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: The Dorsey Press.

Hunter, J.E., Schmidt, F.L. & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Ironson, G.H. (1982). Use of chi square and latent trait approaches for detecting item bias. In R.A. Beck (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore: The John Hopkins U. Press.

Ironson, G.H., & Subkoviak, M.J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Lord, F.M. (1980). Applicants of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Assoc., Publishers.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. (1970). Statistical package for the social sciences (2nd ed.). New York: McGraw Hill.

References (con't)

Shepard, L.A. (1982). Definitions of bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9-30). Baltimore: The John Hopkins U. Press.

Tittle, C.K. (1982). Use of judgemental methods in item bias studies. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore: The John Hopkins U. Press.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: John Wiley & Sons.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user guide: Logist 5, version 1.0 (Computer program manual). Princeton, NJ: Educational Testing Service.

Author Notes

I would like to thank Mary Anne Lahey and John G. Veres for providing the data set which was used in the study. I would also like to thank Rick L. Garvin for his invaluable help in the preparation and analysis of the data.

* * *

Item Bias Detection Methods for Small Samples

Peggy Giffin, Psychological Services, Inc.
Glendale, California

Several statistical procedures have been put forth for the purpose of identifying biased items in a test. The delta method (Angoff, 1982) and the item characteristic curve method (Ironson, 1982) are discussed in more detail in other papers in this symposium. A third method, the chi-square method (Scheuneman, 1979) will be introduced here.

If the selection of which statistical method to use were to be based entirely upon practical considerations, most developers of employment selection tests would favor the delta method for its relative ease of computation and its applicability to small samples. The item characteristic curve (ICC) method, which requires complex and expensive computations and samples of 1500 per comparison group, is out of the question for the development of most employment selection tests where funds are limited and samples are very often under 100, frequently under 50 for all groups combined. The chi-square method lies between the other two methods in ease of computation and sample size, requiring somewhat more effort than the delta method, but much less than the ICC method to compute, and sample sizes of approximately 50 to 90 for each race or sex group to be compared. Thus it appears that the delta method is the most practical for use in many employment testing settings.

If results obtained by the various methods were consistent, restriction of certain testing settings to one method would not be a problem, but this is not the case. Numerous studies have found that the three methods, although showing greater than chance agreement, still differ considerably on the number of items and the specific items identified as biased. The chi-square and ICC methods agree most closely. Of the three methods, the delta method is generally considered to be the least effective in identifying biased items.

This leaves developers of tests with small samples who desire to use item bias statistics with a serious problem: the one method which can be applied to small samples is the least effective method. This quandary led to the investigation of the chi-square method for modifcations that would render this more powerful method practical for use with small samples.

The chi-square method defines an item as biased if the probability of a correct response differs across comparison groups (e.g., race or sex) for people of the same ability level (inferred from total score). For each item a six (or more) cell matrix is developed crossing three (or more) ability levels with the comparison groups (e.g., black and white or male and female).

A chi-square-like statistic is computed using the traditional chi-square formula:

$$\text{Chi-Square} = \text{sigma} \frac{(O - E)^2}{E}$$

where the observed values are the number in each cell getting the item right and the expected values are based upon the proportion of the total ability group getting the item right. (Note: this statistic is not a true chi-square and cannot be tested for significance using chi-square tables. See Scheuneman, 1979 for more details of calculation.)

Two factors influence the minimum sample size requirement: the requirement of at least three ability levels, and the requirement of an expected value of at least 5 for each cell. Two modifications to the traditional chi-square method were designed, each violating one of these requirements.

Chi-Square 2 - The first modification was to compare only two ability levels with the cut point at the total group median. This enabled all cells in most computations to achieve an expected value of 5.

Chi-Square 5 - The second modification was to ignore the requirement of an expected value of 5 per cell. Five ability levels were specified, with cuts at the quintiles. So many ability groups virtually insured that at least some cells would have expected frequencies far less than 5.

The two modifications were compared using both the Monte-Carlo and "real" data to the delta method.

In a Monte-Carlo study, item responses were randomly generated to a 30-item test for groups of 25 subjects each. Bias was introduced into the responses for Group B, yielding lower observed scores but equal true

scores relative to Group A. Of over 1,000 iterations of data generated and application of the three item bias detection methods to the data, the chi-square 2 method performed the best, closely followed by the chi-square 5 method. The delta method was poorest at identifying biased items.

In order to assess the effect of these item bias detection methods on validity of tests in an employment selection setting, these three methods were applied to data from a criterion-related validation study of language skills tests for nursing staff at a large hospital. The criterion in the study was a work sample involving reading, writing, speaking, and listening tasks in a hospital context. Sample sizes were 24 and 29 for the two comparison groups. Each of the two predictor tests (vocabulary tests) from the study was rescored for each of the three item bias detection methods, eliminating items identified as biased by that method. The revised test scores were then correlated with the original criterion. For one of the two predictor tests there was no significant difference between the validities of the original scoring and the bias-free rescores for any of the three item bias detection methods. For the second test, however, the chi-square 2 and chi-square 5 rescores each correlated significantly higher with the criterion than did the original score. The delta rescore showed no change in validity from the original score.

On the basis of these findings, it is concluded that while all three of the methods investigated identify biased items at better than chance levels with N=25, the chi-square modifications perform better than the delta method in this sample range, without reducing, and sometimes with the effect of increasing, validity. When item bias investigation is contemplated for small sample sizes, the two chi-square modification methods are recommended over the delta method.

References

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore and London: The Johns Hopkins University Press.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore and London: The Johns Hopkins University Press.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

* * *

PHYSICAL PERFORMANCE TESTING FOR FIREFIGHTERS (Paper Session)

Chair: Alfredia Boyd, South Carolina Division of Human Resource Management

Discussant: Cassandra K. Scherer, Milwaukee Fire and Police Commission

Physical Test for Firefighter

Ester K. Juni, New York City Department of Personnel
Captain George R. Layng, New York City Fire Department

New York City is, unquestionably, the high-rise capital of the world. In its five boroughs, there are 5,000 buildings that range from 100 feet to 1,350 feet. Understandably, with this number of high-rise buildings, New York's Fire Department has had the greatest high-rise fire experience. New York City is also considered the melting pot of various commercial occupancies, the port of New York merits noting. The port of New York, largest in the world, has 650 miles of usable waterfront. This is serviced by more than 500 piers, which are capable of handling 15,000 ocean-going vessels annually.

The most recent Fire Department annual statistics (1982) reflect the following incidents and deaths throughout New York City:

A.	Total Fires	111,799
	Total Emergencies	77,132
	False Alarms	152,147
	Total Incidents	341,078
B.	Civilian Deaths	233
	Firefighter Deaths	4
	Total Deaths	237

When one considers the above facts and statistics, it becomes apparent that a New York Firefighter must be well qualified, both mentally and physically, to perform at a level demanded by the City's various physical complexities and magnitude.

The present firefighting force is equivalent to a small quasi-military army. The morale of this army is extremely sensitive to required entrance standards, since any delution of accepted standards is considered a threat to both the safety of the citizens they are sworn to protect, as well as to their own personal safety.

Considering the above facts, coupled with the acknowledgement that federal courts carefully scrutinize every exam, with the intention of controlling possible adverse impact, the responsible Agencies (Fire Department) combined forces in an effort to produce a firefighter physical examination that would follow federal uniform guidelines, yet still provide the best firefighting force possible. In consideration of

all the above factors, the following job-related firefighter physical examination was constructed.

The New York City firefighter physical exam is a timed, competitive test, which is divided into three major segments.

Part I is an engine simulation. The candidate drags a 3½ hose 150 feet, then places a folded hose on to his/her shoulder, carrying it up three flights of stairs and a distance of 85 feet. (S)he places this hose on a bench, and then pulls a length of weighted 50 foot hose in through a simulated window.

Part II is a mandatory rest period. From the moment the hose pull segment of the test is completed, the candidate begins a 100 second mandatory rest period. During this rest, the candidate must walk down three flights of stairs and proceed to the start of Part III. The distance is approximately 370 feet. At the end of 100 seconds, the signal for Part III begins, regardless of whether the candidate is there and ready to proceed.

Part III is the ladder simulation. The candidate begins by scaling a 4½ foot wall and immediately proceeds 75 feet to the ladder raise. After raising the 20 foot ladder, (s)he ascends and descends another ladder, approximately 10'. At the foot of this ladder, the candidate picks up a 15 pound weight and ascends three flights of stairs. At the top of the stairs, (s)he puts down this weight and performs a forcible entry simulation, hitting a 77 pound weighted tire down a 12½ foot metal topped table with an eight pound mall. This is followed by a crawl through a "U" shaped tunnel, approximately 25'. The final event is a dummy drag, pulling a 145 pound articulated dummy around the table used in forcible entry.

Throughout the entire test, the candidate wears a weighted vest, the weight being concentrated on the hips and lower back, and a Scott air tank on his/her back. The total weight, 40 pounds, simulates the weight of the gear worn by firefighters on the job.

This test is of short duration, requiring great energy and endurance.

* * *

INVITED SPEAKER OF THE WESTERN REGION INTERGOVERNMENTAL PERSONNEL ASSESSMENT COUNCIL (WRIPAC)

Chair: Terry S. McKinney, City of Phoenix, WRIPAC President

Historical and Future Perspectives on Assessment Centers

Cabot L. Jaffee, President, Assessment Designs International
Orlando, Florida

The use of simulations for evaluating purposes is most often traced to the early work conducted by the U.S. Office of Strategic Services in World War II (Assessment of Men, 1948).

In 1956, the American Telephone and Telegraph Company initiated a longitudinal research study focusing on 422 young managers. The purpose of this study was to determine the value of a series of evaluation devices, some of which were simulations, in predicting the progress made by these individuals over time.

Nineteen-fifty-eight witnessed the introduction of the first operational assessment center in American industry by Michigan Bell, which implemented the concept for elevating craft employees in first-level supervisory positions. This effort also was characterized by the total reliance upon managers as evaluators/assessors as opposed to psychologists (the use of internal organizational personnel as evaluators/assessors has since become the most frequent choice by organizations using the assessment center concept).

In the 1960s, the use of the concept grew gradually with such organizations as IBM, General Electric, and Standard Oil of Ohio beginning to use it. Toward the end of the decade, federal government agencies began to implement the concept within their organizations. During this same period, scientific studies started appearing as well as articles meant for practitioners (e.g., Jaffee, 1966; Jaffee, 1967; Meyer, 1968, Bray and Grant, 1966; Bray and Campbell, 1968).

The 1970s are associated with a significant upswing in its usage, primarily because of fair employment practice considerations and organizations' greater commitment to employee development, both of which are viewed as being served by the assessment center process. In the decades of the 1970s and into the 1980s more and more organizations began seeing the assessment center method as having broader applications than simply for selections. Career planning and the overall development of employees were seen as areas that could benefit from the assessment center concept (Bender, 1973; Jaffee, Frank and Rollins, 1976). Most recent estimates of the number of organizations which have used the concept have been as large as 2000 (Parker, 1980).

Elements of the Process

While assessment center applications may vary from one organization to another, or even within the same organization, there are certain standardized elements that are found in all quality assessment centers. Given that there are many differences in application, it's probably safe to say that all assessment centers have the following individuals that make it work.

Participants: are those individuals being assessed. They may be referred to as candidates or assessees. All probably went through some form of prescreening which can range from panel interviews to a review of recent past performance. Once at the assessment center, each participant is asked to complete a predetermined schedule of simulation exercises.

Exercises usually number from three to five, and along with beginning orientation and breaks, a participant is usually involved in actual assessing from one to four days. Participants may also be asked to prepare written reports, summaries, or outlines which represent actual or proposed future actions to be taken. Upon completion of all exercises by all participants--there are usually from six to 18 participants at each center--a group debriefing session is held. The purpose of the debriefing is to answer participants' questions about the process, to summarize the events of the last day and a half, and to provide directions for any activities to follow. Debriefing is not for the providing of feedback on performance, which can be given only after additional work completed by the assessors. With the exception of participants from "outside the organization" who are being assessed for initial hire, most assessment center participants receive verbal feedback as the final segment of the center. In addition, a more detailed final report is usually presented at a later date. Both oral feedback, and the more indepth final reports, stress observed job-relevant strengths and weaknesses, as documented by a team of assessors. The information is considered confidential, and as such, it only provided to the participant and those selected few with a legitimate need for the information.

Assessors: are those doing the actual observation of the participants as they go through the simulation exercises. In most instances, they are staff members of the organization (internal assessors) who are trained as observers and released from their regular day-to-day responsibilities to function as assessors. Another source, although not used as often as internal assessors, is people from outside the organization (external assessors), who have been contracted with to provide the assessing function. Finally, there are cases where, for various reasons, a combination of internal and external assessors is teamed together.

With the primary purpose of an assessment center being to collect quality information, the proficiency of the observer, then, naturally becomes paramount. To ensure this critical element, assessors complete rigorous, certified training programs.

Assessors' activities can briefly be described as observing, classifying, and scoring of job relevant behaviors which are demonstrated by participants as they experience simulation exercises. Assessors usually work in the "teams" of three. While individually they assess one participant at a time, the team usually assesses six participants over the course of the center (typically three to five days). Assessors must individually complete written reports for every exercise observed, usually six to eight reports per center. As a team of three, assessors must then meet to reach consensus on all scores for each skill for each participant that they have observed.

Role Players: Since an assessment center is a simulation of actual job situations, it usually needs the involvement of more than just the participating assessee. In other words, on the job, one frequently interacts with others. Therefore, during the assessment center, one should expect to interact with others--the role players.

The purpose of the role players as they carry out the interactions in the exercise is to enhance realism. Their primary responsibility, though, is to ensure that participant #1, participant #10, and all the other participants encounter the same type of standardized interactions. As such, the role players fulfill a very important function in the assessment center by ensuring that each interaction is consistent for each participant.

For the role players to carry out their responsibilities, they must have some prior guidance and training. The guidance they receive is in the form of thoroughly prepared instructions, or guides, which explain the actions they are to take during the exercise interactions. The training they receive focuses upon learning the exercise materials and participating in actual practice exercise sessions. With this guidance and training, the role players can interact with participants in a prescribed, real-life manner.

Administrator: Basically, the function of the administrator is to ensure that consistency or standardization occurs during each assessment center. In many respects, the administrator provides a quality control function to make each assessment center the same. The responsibilities of the administrator fall into three general categories: 1) pre-assessment center activities; 2) assessment center activities; and 3) postassessment center activities. The pre-assessment center activities consist of both information sharing and program coordination. Information sharing should begin immediately after the decision is made to conduct an assessment center. Once this information has been shared, program coordination begins. The administrator schedules the role player and observer training programs while also handling the logistics for the actual assessment center.

The second category of administrator responsibilities, assessment center activities, directly impacts on the participant. To begin the assessment center, the administrator holds a program orientation, or briefing session with everyone attending the program. At this session, participants are informed about the skills that will be evaluated and the exercises in which they will participate. Additionally, instructions about the general guidelines which must be followed are provided with a schedule of activities. After the briefing session, the actual exercises are ready to begin.

The administrator's postassessment responsibilities consist primarily of "wrap-up" activities. The administrator collects all program materials and from these materials starts and maintains data files. He/she serves as a liaison to the people who have access to and will use the file information. Finally, the administrator may conduct an evaluation of the assessment center program.

Simulation Exercises: The exact exercises that a participant goes through in an assessment center vary from one center to another, based on the job analysis of the focal position(s). Please keep in mind that

exercises usually consist of a specified amount of time for material review; and then some interaction with one or more role players in a meeting, presentation, or interview format. It's during this interaction phase of each exercise that the participant is observed by a trained assessor.

Traditionally, exercises have included group discussions, business games and in-baskets, perhaps more than any other types of exercises, but this is certainly changing drastically. So the past is by now solidly based on good research data, perhaps 75,000 people per year going through assessment centers of all levels and for many, many purposes.

There have been, and will continue to be, greater use of the assessment center idea of competency based testing. Professionals of all kinds are being assessed using simulations. Physicians, dentists, attorneys, are being placed under simulated conditions in order to better evaluate their job-related performance. Teachers, chaplains, police officers, military officers, and obviously all the uses most of us are already familiar with, speak to an ever widening application of the technology. But what will happen in five to ten years in regards to industrial applications? I see two major directions. One, the changes that new technology will bring to assessment centers; and two, the greater diversity within industry as to the goals of the process. In other words, assessment centers will be done differently and used differently than they are presently. In the following examples, I will try to point out the directions somewhat more specifically.

Job Analysis

The traditional procedures of interviews and questionnaire analysis of frequent and important tasks, will be streamlined considerably. We already are doing some interviews over the telephone and electronically analyzing questionnaire results as they are sent in without any need for a "human calculation." Eventually job analysis will likely become a totally automated process in which you might go from beginning interview to final report on very large samples within perhaps a day.

Assessor Training

The technological advances described above will make assessor training quite different in the next ten years. We have already developed, using a combination of video and computers, a self-training module for assessors in which highly specific check lists, video vignettes and a carefully programmed workbook, allow assessors to be trained in their own offices, including a final check on their abilities.

Conducting Assessment Centers

The same technology impacting on assessor training will also change the nature of assessment centers. We are conducting assessment centers in which the exercises are conducted at remote locations, videotaped and sent to a central location where teams of trained

assessors can process, extremely cost effectively, the performance of the candidates. Electronic mail and teleconferencing rooms will allow for this to operate with greater interaction between assessor and candidate. A number of other changes center around the use of videotaped interactions with applied multiple choice tests to evaluate large numbers of people when an assessment center, by traditional definition, is not feasible.

After the Assessment Center

Computers can now write final reports, provide vocational guidance, perform career development activities, and provide feedback about assessment center performance. All these things will continue to be done by people for awhile, but gradually, as the need for greater efficiency continues, the computer will take over many of these functions, and by that time, people will not be disturbed by the lack of human contact for the transmission of such important personal information. In fact, recent articles seem to support that even in areas of psychotherapy, a machine has some comforting objectivity even beyond the trained professional.

In summary, we see some strong forces at work which will influence the future of assessment centers and not all those forces will work in a common direction. Clearly, the need to do things more efficiently will lead to greater use of technological advances to make the assessment center less costly, and therefore, potentially valuable for lower level positions for which present assessment centers could not be cost justified. On the other hand, this same technology will allow for better simulations of work environments, and that clearly will be a trend. A better simulator, modeled after the aircraft simulator, will definitely be an idea whose time has come over the next five years. This will increase accuracy of selections as the test becomes more like the job, and for the more important positions, will be more likely to be used. In terms of usage, assessment in colleges and universities will grow, as a greater realization of cost effectiveness will place the assessment center at the widest part of the people funnel and create the need for more sharing between companies in the development of people resources.

* * *

BIODATA AS AN ALTERNATIVE SELECTION TECHNIQUE: AN EXTENSIVE EVALUATION (Symposium)

Chair: Glenda K. Corcione, New York State Department of Civil Service

Presenter: Glenda K. Corcione

Introduction/Overview

The New York State Department of Civil Service Biodata Project is being conducted on a research basis to determine if a biographical instrument can be developed to accurately predict performance and tenure of Mental Hygiene Therapy Aids and Trainees. I am responsible for administering the grant awarded to fund the study, as well as overall coordination of all activities related to the project. Participants include the Department of Civil Service, two of New York State's line agencies - the Office of Mental Health and the Office of Mental Retardation and Development Disabilities and our consultant, Dr. Robert Means.

Bob Means co-founded the consulting firm OXICON in California recently acquired by McGraw-Hill Publishing Company, Inc. As an expert in the field of biodata, Bob is assisting New York State in the development, administration, and validation of a biographical questionnaire for the title Mental Hygiene Therapy Aide Trainee.

Background

The Mental Hygiene Therapy Aide title comprises the largest number of incumbents in any given title in New York State - over 2,000 trainees and 20,000 aides. After successfully completing a one year traineeship, trainees are automatically advanced to the journey level status of Mental Hygiene Therapy Aides. These positions are located in 48 mental health and mental retardation facilities statewide.

Our current selection process for the Trainee position requires that an individual can read, write and speak English, as well as compete on the written examination which tests for understanding the care of the mentally ill and "disabled." The salary for this entry-level position is $13,917 which is, for most parts of the State, a very attractive entry-level salary to many individuals who have no specialized education or experience. However, once they get in and discover that they have to change diapers on adults, ward off abusive behavior and spend months teaching someone to put a spoon in his/her mouth, many come to realize "Hey, this is not for me!"

This realization has led to significant performance and tenure problems in the first year after hire which in turn, translates to high costs in recruitment, training and counseling, not to mention the decreased quality of care to patients, and overwhelming cost to taxpayers.

In calendar year 1984, turnover at the trainee level was 37% (921 employees). Of those 921, 44% (405) were terminated during their probationary period. The remaining 56% (516) resigned. Cost of this turnover -- $2,436,045. These statistics are particularly alarming when one considers that Mental Hygiene Therapy Aides and Trainees are key providers of primary patient care.

In addition to the cost factor, morale among current employees is low. While trainees are in classroom training, an unreasonable burden is placed on current staff who are forced to care for more patients than normally planned for, and who are forced to work overtime when coverage on the next shift is insufficient. This develops into a vicious cycle, causing absenteeism due to illness and fatigue, which causes more overtime and possibly a low level of performance for those remaining Aides and Trainees.

Why Biodata?

Previous studies by the Department of Mental Hygiene raised questions about the predictability of the current written test and the Department of Civil Service staff began exploring various alternative selection mechanisms. Oral test and assessment centers were too expensive for the volume of applicants we get applying for the Trainee position. With 26,000 applicants per year a training and experience exam was also not practicable.

Department staff had become familiar with biodata from a previous IPMAAC conference, and persuaded department management to consider exploring biodata as a possible selection approach for one of our problem titles. We conducted a literature search and determined that biodata might meet New York State's selection needs. In many cases, biodata has been found to be highly predictive of performance, highly predictive of tenure, without adverse impact, and once developed, cost effective.

Although never tried in New York State government before, biodata has been successfully used in the past for other populous job titles. All this considered, biodata seemed a viable option to pursue on a research basis.

Issues to Consider

The issues we faced in this project are primarily due to the fact that we were working with two large agencies with multiple facilities statewide, two unions and professional and paraprofessional staff. Four critical considerations include: the level of central personnel agency support; who has the technical expertise; the level of agency support where the title exists; and the financial constraints imposed on your organization.

Internal Political Barriers

Two major internal political barriers exist - contracting with a private consultant and obtaining the support of the unions. If you are a public jurisdiction contracting with a private consultant, sufficient time should be allowed for the contract to go through the appropriate channels. Be prepared to overcome a series of obstacles. Making your organization's needs mesh with the consultant's needs, may require contract language compromises at every level of review. If you are working with a consultant from another state, make sure you have made him/her aware of state law in terms of testing and selecting people. There are variations from state to state.

If your state is unionized, support of the unions is critical. Timing is everything! What else is going on at the time you are asking for their support and cooperation? The sun was not shining on us when we were ready to approach the unions. They had an annual major convention planned, election of officers scheduled, a three-year contract being negotiated and labor/management problems in individual facilities. Needless to say, biodata was not uppermost in their minds.

If you are not unionized, what mechanisms are available for soliciting cooperation from the involved employees?

Logistics

Regardless of the size of your organization, a centralized coordinator will need a scheduling technique such as PERT to plan logistics through three major development phases. (Interviewing, item development and printing). If you are a decentralized organization, you will want, in addition to the centralized coordinator, a contact or coordinator in each facility or agency.

Administration

This is the most critical phase of the project. The more people who participate, the better your chance for success. Additionally and more importantly, for each employee who completes a questionnaire, you need a supervisor to complete an evaluation of that employee in order to establish a "match."

The current project is such a massive undertaking for New York State - more than 22,000 people participating - that we attempted to set up a network of communication. I mentioned earlier the general meeting of the personnel officers. Although this was a good first step, in retrospect we should have done more. Our description of the project and our instruction for the forms were understood more by some than by others. The slightest bit of confusion was magnified ten-fold by the time it got to the field. Perhaps more publicity up front in facility newsletters would have helped.

At any rate, you can initially expect a lot of phone calls with questions regarding the forms but later, there will be a need to follow-up with those that you have not heard from to ensure that distribution has occurred. Personnel Offices tend to be overworked and understaffed, and a little prodding may be necessary.

It was during the administration stage that NYS experienced its two most serious problems:

1. The unions began to resist the project. They were very concerned about the inclusion of social security numbers on the response sheets. They felt this was an invasion of privacy and had the potential of being used against them later.

2. Some of the supervisors, also unionized, objected to the performance evaluation forms. We would hear comments like, "Will this form be compared to the standard state performance evaluation form I've already completed? Do I have to put my employees' SS# down?"

   A. The rumors began to fly. By the time those misconceptions got back to me, I found out that this project had many interesting purposes. We were collecting information on people to develop a layoff roster; we were using the information generated by less than satisfactory employees as a club to get rid of them later and best of all, we went into the profit-making business by selling their names and SS#s to other organizations.

   B. Many meetings and telephone calls later, the project continued on course. We are now averaging about a 33% response rate. The % of employee/supervisor matches is yet to be determined. Bob has already given you a preliminary analysis, based on what we have received so far.

Lessons Learned

Based on our experiences to date, the key to success is involving all relevant parties on an ongoing basis, but there are three points I would like to emphasize.

1. Your expectations should be made crystal clear to the field. We never considered that filling out performance evaluation forms would be a problem for supervisors. Isn't that part of their responsibility? The unions did not agree and forced us to make a statement as to the voluntary or mandatory nature of supervisor participation - which brings me to my next point.

2. Involve the unions as much as possible as early as possible. Do not let a sleeping dog lie. If you do not receive an opinion or hear from the union regarding your project, seek one. Make sure they have a total understanding of what will help their membership.

3. Plan as far in advance as you possibly can. As the project progresses, you will be putting out a lot of brush fires.

Sound Like a Lot of Work? It is, Is it Worth it? That Remains to Be Seen?

We are hopeful that the data will show biodata to be a worthwhile alternative selection technique. If it does, the work we are expending up front will save us years of extra work later, not to mention saving the state millions of dollars. We would have a selection instrument for the trainee title that predicts performance and tenure and which would not have to be redeveloped, only revalidated.

If the process works for us as it has worked for others, we would not expect to be dragged into costly court battles challenging our selection instrument because it will not have adverse impact.

Can It Be Used In A Public Jurisdiction Within The Constraints Of A Merit System? We Believe So!

It is competitive. It can be ranked or zone scored. It is criterion-validated and previous experience has shown that it is designed to not have adverse impact. The New York State study is tentatively scheduled for completion on December 31, 1985. At that point in time a final report and recommendations will be made to the Commissioner of the New York State Department of Civil Service.

* * *

DEVELOPMENT, VALIDATION, AND USE OF A STATE POLICE ENTRANCE EXAM IN A CONSENT DECREE ENVIRONMENT (Symposium)

Chair: Elizabeth H. Mackall, Maryland State Police

Participants: Colonel W. T. Travers, Maryland State Police; and Richard H. McKillip, Psychological Services, Inc.

The Equal Employment Opportunity Act, signed into law March 24, 1972, extended the Title VII provisions of the 1964 Civil Rights Act to state and local governments and empowered the office of the United States Attorney General to bring civil action in federal district court against alleged violators. The Maryland State Police was the first state police force sued by the Justice Department under the powers given to it by the new act.

Our selection procedures at the time were not much different from those found in other state and local jurisdictions. In common with most other police agencies we barred women from enforcement in employment positions. And although the representation of blacks on our sworn force was not quite four percent, this utilization compared favorably with other major jurisdictions, including those with much larger minority populations.

In the year-long discussions with the justice department that preceded the signing of the consent decree, our written test was a key issue.

Following the semi-annual administration of this test, the department of personnel would prepare a certified list of eligibles composed of all those with test scores of 70 or higher, and would then issue this list to us so that we might conduct the remaining selection procedures. As was typical with many other written tests then in use throughout the country, the examination had an adverse impact on blacks, but no evidence had been gathered in support of its validity. Since it was

administered according to merit system rule, moreover, the department of personnel could not intervene to modify cutoff scores nor in any way minimize its impact on minorities.

Consequently, when the consent decree was signed finally in January, 1974, it explicitly prohibited use of the existing test, or any other written test unless

> "... such test has no disproportionate adverse impact upon blacks or has been validated in accordance with the EEOC's (then) current Uniform Guidelines on Employee Selection Procedures."

The goals set forth in the consent decree pertained initially to utilization of blacks in the overall sworn force. In 1979 this portion of the decree was modified to provide for specific hiring goals for both blacks and females at the entry level.

With hiring opened to women for the first time, with an intensive minority recruitment program established, and with the loss of the written test as a screening hurdle, the applicant pool more than doubled within a short period of time.

Several efforts were made to re-establish a written test. First among these was an attempt to validate the existing examination by correlating the exam scores of current employees with available supervisory ratings. This study immediately ran into problems and had to be discarded. A second effort at re-establishing a written test was made in collaboration with the educational testing service which was launching a multi-jurisdictional testing project for the IPMA and IACP jointly. The validity results were available in 1976, but, to our dismay, suggested that this examination would have an adverse impact exceeding that of the test that had precipitated the consent decree. We accepted an invitation to participate in a second multijurisdictional testing project which was being funded by LEAA under the direction of John Furcon from the human resources center at the University of Chicago. Unlike the IACP/IPMA test, which focused solely on intellectual abilities, this new test was to measure several other skills, abilities and personality characteristics not thought to differ by race. Unfortunately, pilot testing in our agency and initial administrations from other jurisdictions indicated that adverse impact was still present. It was also apparent that the test would be both too costly and too cumbersome for us to administer on an on-going basis.

By 1980, then, it had become clear that if we wanted a test, it would have to be tailored to our needs and constraints. It had become equally clear by this time that we had a critical need for a test. Our applicant pool had more than doubled, and the lack of an efficient screening device meant that we were needlessly spending vast sums of time and money on applicant processing.

With the exception of the final processing step, the oral interview, all processing procedures were administered on a pass/fail basis with cutoff scores set at the minimal acceptable level. Under these conditions more than half of all applicants were typically passed all the way through from the first step in processing, the physical agility demonstration, to the last step, the oral interview. The middle step in processing, the background investigation and polygraph, is extremely expensive to administer because it is so time-consuming.

It was particularly frustrating to conduct backgrounds and polygraphs on applicants who were clearly functionally illiterate -- and they came to us in droves because it was well-known that we had no test -- and then have to pass them through to the next phase because they possessed a high school diploma and ipso facto were qualified according to the terms of our consent decree.

Processing inefficiency was only one of the problems necessitating re-introduction of a written test; measurement unreliability was the other. Although the oral interview had been restructured entirely so that it had a fair degree of validity with respect to predicting subsequent performance in the academy and on the job, its reliability simply was not strong enough to ensure that all candidates selected would be successful.

Even though terminations due to poor performance were limited in number to between one and three candidates per class hired, they were expensive, costing an estimated $20,000 per candidate in training and equipment alone. Furthermore, since the vacancy created by a termination might go unfilled for a year or more, we could ill-afford mistakes in our selection process.

Despite the fact that the need for a written test had become critical, we were unable to get funds budgeted by the legislature for outside assistance for its development. Thus, with no apparent recourse left, we decided to proceed on our own.

After discussing a variety of strategies, the project got underway finally in the fall of 1981. The training academy was selected as the focus for the design and validation of the test because it offered a number of advantages. First, the content and instructional domain were finite, stable and measureable. Therefore, a content valid work sample approach could be employed that would obviate the need for any hypotheses about required knowledge, skills or abilities. Second, reliable and objective data in the form of test marks and coursework averages would be available within a relatively short period of time for a predictive criterion related validity study. Third, data collection for both test development and validation would be relatively inexpensive and unobtrusive.

In order to use training performance as a criterion in test validation, the Uniform Guidelines require that the "relevance of the training be shown either through a comparison of the content of the training program

with the relevance of important work behaviors on the job, or through a demonstration of the relationship between measures of performance in training and measures of job performance".

Since the entire purpose of the training academy is to prepare candidates for the road patrol position, demonstrating the correspondence between the content of the academy and the content of the job was very straightforward.

From available records we were certain that there was also a close correspondence between performance in the academy and performance on the job. To demonstrate this statistically we brought in a group of field supervisors to serve as subject matter experts and, using the job analysis data gathered earlier, to construct a job performance rating instrument. We then identified 160 graduates from three consecutive academy classes, gathered the data pertaining to their academy classes, gathered the data pertaining to their academy performance, and sent out a special order to the field requiring that first and second line supervisors independently rate the job performance of the targeted employees. To avoid contaminating the results we did not disclose the purpose of the project (all those involved thought we were merely piloting a new evaluation system). In addition, to minimize the tendency towards leniency that typically plagues our supervisory ratings, we asked that raters not discuss the ratings with anyone, including the rated employee.

We received completed performance ratings on 143 of the graduates, converted the ratings to numerical scores, and then correlated them with final academy averages. The coefficient obtained for the entire sample was .50. When corrected for unreliability in the criterion (and we anticipated a degree of unreliability since we had instituted the ratings without training or advance warning), the coefficient rose to .57. The coefficients for blacks and whites are similar but they exhibit a degree of depression due to restriction in range. The coefficient for females, although smaller, is not significantly different from that for males. The fluctuations in inter-rater reliability and correlation coefficients for females were probably due to the small sample size.

To arrive at the best estimate of the true relationship between academy performance and job performance we standardized the ratings for each class separately. This gave us a measure of job performance unaffected by length of tenure. The correlation coefficient resulting from this run, when corrected for unreliability in the criterion, was .62. At this point we were satisfied that we had convincing evidence that the relationship between the academy and the job was sufficiently strong to justify using the training academy as the focus for test design and validation.

We intended to use a content validity approach in designing the test. The Uniform Guidelines state that "a selection procedure can be supported by a content validity strategy to the extent that it is a

representative sample of the content of the job." The Guidelines in turn define the content of the job as the set of observable work behaviors and work products. In a training program such as our academy the work behaviors are largely unobservable mental processes -- how the student goes about learning the material presented. The work products however -- the reports, tests, in-class presentations --are observable, and serve as measures of performance of the work behaviors. Consequently, our working premise was that the content validity of the test could be supported by the degree to which the test was able to simulate the training situation through a representative sampling of the subject matter presented and the work products required.

To gather the necessary information we constructed a questionnaire to be completed by the instructors responsible for each of the 16 different academic subjects taught in the academy. The questionnaire asked for information on format of instruction, reading materials assigned, sources of test items, test item types, and skills tested. The responses to each questionnaire were then weighted according to that subject's percentage contribution in calculating a student's overall academic average (40 tests are used in an essentially unweighted combination to arrive at a student's overall academic average. There are eight tests devoted to the subject of motor vehicle law. Thus the weight given to the Motor vehicle law questionnaire was 8/40 or .20). These weighted data were then compiled to provide an overall summary analysis of the instructional domain.

Although we thought the test content valid, we had planned all along to gather predictive criterion related validity evidence before approaching the Justice Department. To do this we administered the test to applicants to two consecutive academy classes.

Two and a half weeks prior to the scheduled test dates all active applicants were sent a study handbook along with information about what to expect on the test and how to prepare themselves for it. In keeping with our Consent Decree, the scores from the test were to be used for research purposes only, but we attempted to obscure this fact from applicants so that their motivation to do well would not be dampered. On the morning of the test date applicants first viewed videotaped lectures, following along and taking notes in their study handbooks. A one hour study period followed the videotapes. Although the purpose of the study period was to allow a candidate time to try to integrate the material presented in the videotapes with the information from the handbook, they were permitted to use this time as they saw fit and many chose to sleep or socialize. At the end of the study period all handbooks were collected and the examination was administered. There was no time limit but most finished within an hour and a half.

A total of 329 applicants were administered the test. The range in scores was good with an overall standard deviation of around 12 points. The internal consistency as measured by Kuder-Richardson Formula 20 was entirely satisfactory. There was, however, clear evidence of adverse impact, with blacks scoring about one standard deviation below

whites. Score distributions for males and females were quite similar in all respects. The differences between the two classes are attributable to differing proportions of the two race groups represented in the examination pool.

The predictive-criterion related study took over a year to complete because we had to wait six months for each class to graduate before correlating the exam scores with their academic grades and averages in the Academy.

The validation samples for each class were quite small, consisting of only those taking the exam prior to selection and subsequently graduating from the Academy. There were 36 in the 82nd class sample, 42 in the 83rd class sample.

The correlation coefficients between the examination and major coursework areas as well as final academic average were extremely high and stable. The differences in the obtained coefficients for the two samples were due largely to greater restriction in range on the exam scores for the 83rd academy class. When corrected for restriction in range the coefficients were nearly identical. The results were analyzed further to determine if there were any significant differences between residuals, slopes and intercepts for the two classes and the regression data for the two classes was remarkable similar.

We were quite gratified by the initial validation results, but since this was our first venture into the field of test validation, before approaching the Justice Department we contracted with Psychological Services to review and critique our preliminary technical report in detail.

Stephen Bemis and Dick McKillip reviewed the report with a fine tooth comb and gave us a number of very valuable criticisms. They recommended that we conduct a test fairness analysis using the Cleary method and a utility analysis using the Schmidt, Hunter method. Most importantly they suggested a method for using the examination that would enable us to meet our Consent Decree goals while still achieving our other major objective of reducing processing costs. It took another four months to implement their recommendations and to rewrite the technical report.

* * *

CAREER DEVELOPMENT ASSESSMENT CENTERS IN PUBLIC AGENCIES (symposium)

Chair and Participant: Karen Coffee, California State Personnel Board

Participants: Dennis Joiner, Dennis Joiner and Associates; and Stephen Boles, San Mateo County, California

DIFFERENCES BETWEEN SELECTION AND DEVELOPMENT IN ASSESSMENT CENTER PURPOSES

Control/Activity Area	Selection	Development
Job Analysis	Critical.	Important to degree that development is fitted to particular job.
Selecting dimensions	Critical; limited number	Important. More dimensions might be used than in selection.
Candidate expectancies	Important; candidate should know the purpose is selection.	Important that candidates know selection is not involved; development is.
Review alternate measures	Useful as efficiency check; choices determined empirically.	Many experiences/ measures might qualify. Professional judgement is called for.
Check candidate performance · across dimensions · by dimensions	Critical as check on scale effectiveness and/or rater performance.	Useful as basis for describing candidate performance; agreements between raters need not be high.
Check interrater agreement	Critical; interrater agreement by dimensions and total score is important.	We can tolerate some variability among raters since prediction is not involved.
Internal consistency of report	Important.	Since contradictions might be basis for fruitful development discussions, no major concern. A written report might be omitted.
Feedback	Critical; control on timing and form is necessary; limited in scope.	Can be intermittent/ informal; could be conducted during assessment; could be group; could use video.
Interventions	Time limited and fairly specific; geared to specific job. Often group and not tailored to a single individual.	Not predictable as to content until job target is fixed; requires good follow-through.

DIFFERENCES BETWEEN SELECTION AND DEVELOPMENT IN ASSESSMENT CENTER PURPOSES (con't)

Control/Activity Area	Selection	Development
Assessor role	Raters specific behavior using a rating scale or schedule. Judges goodness.	Describes a wide range of behavior and suggests areas of weakness/ strength. Solicits other candidate or participant observations. Encourages discussion of participant behavior.
Assessor skill	Should be familar with limited job so judgement of goodness is enhanced. Should be familiar with rating scale (what goes with what dimension. Should be skillful at ignoring behaviors not relevant to the constructs (dimensions) being measured.	Need not limit his/her contribution to a single job or job family since some participants may not want to be managers. Should be skillful in suggesting a wide range of developmental possibilities for each participant. Should be skillful at identifying resistance.

Material adapted from: Keil, E.C., Assessment Centers: A Guide to Human Resource Management. Reading, Massachusetts: Addison-Wesley Publishing Company, 1981.

* * *

INVITED SPEAKER OF THE PERSONNEL TESTING COUNCIL

Chair: William W. Ruch, Psychological Services, Inc.

Uniform Guidelines on Employee Selection Procedures: A Proposed Alternative

Keith Pyburn, McCalla, Thompson, Pyburn & Ridley,
New Orleans, Louisana

In our society regulators and courts should address issues up front in a direct, rational manner. Lest I be accused of being foolish let me hasten to add that I do not believe this always happens. Frequently the failure to address problems directly and within the bounds of our understanding leads us to wasteful, if not ridiculous, rules of law that are more cumbersome and expensive to administer and follow than could ever have been envisioned.

The example of this failure to address issues directly is the disparate impact theory of employment discrimination as applied to Employee Selection Prodecures and as codified in the Uniform Guidelines on Employee Selection Procedures.

The history of these regulations and the court battles which have grown up around them, is a prime example of costly regulation which has led us in a circle with irrational results the order of the day.

The theory of disparate impact discrimination was first announced by the U.S. Supreme Court in the case of Griggs vs. Duke Power Company. In this case the Court set forth what appeared to be, at the time, relatively simple rules. The Court interpreted Title VII of the Civil Rights Act of 1964 to prohibit not only intentional discrimination, but discrimination which was caused by procedures fair in form, but discriminatory in effect.

In the context of reviewing the use by the defendant of two employment tests and a high school education requirement, the Court announced the principle that any procedures used by an employer which disappropriately excluded minorities or other groups of individuals from employment opportunities were unlawful unless the employer could show the procedures in question were "job related." This principle has been codified in the Uniform Guidelines.

It is interesting the Supreme Court announced this theory of discrimination in the context of a case where, upon careful examination, the court could have ruled for the plaintiff based on an intentional discrimination theory. The facts of the Griggs case demonstrate substantial evidence of intentional discrimination. Specifically the use of the tests which were challenged in the case essentially replaced a per se rule of segregation. The tests were adopted essentially on the day Title VII became applicable to the employer. Further, there was absolutely no study, not even a review of the usefulness of the tests for the employer's operation. It had to be known that the high

school education requirement would disproportionately exclude blacks. Despite this, some blacks met the requirements, and yet were not admitted into the previously segregated white departments for a number of months. From all of these factors the Court could have inferred there was a pattern and practice of intentional discrimination. Instead, the Court adopted the theory of impact discrimination. It has had far reaching implications and has been extemely difficult for the courts to administer.

The disparate impact policy announced in Griggs has been incorporated into the Uniform Guidelines on Employee Selection Procedures. The debates and the compromises that were integrated into the Guidelines is a story in and of itself. Nevertheless the Guidelines certainly directly track the Griggs principle.

Shortly after the Griggs decision, the problems with administering this rule of law began cropping up. There are three general problem areas caused by this rule: 1) The fact the courts are forced to ignore the Guidelines to avoid riduculous results (and arguably sometimes don't); 2) the cost; and 3) the requirement there be a showing of validity which "meets professional standards."

The first problem surfaced almost immediately. An illustrative case is Spurlock v. United Air Lines. Here United Air Lines had a requirement that applicants for the job of pilot would only be considered if they had a college degree. There is nothing in Title VII or in Griggs which asserts high school degrees are somehow different than college degrees. Nevertheless, in this case, without relying on any study, the Court held the college degree requirement for pilots, even though it clearly had a disparate impact, was lawful because pilots had to undergo complicated, rigorous training. The Court concluded the college degree would help these pilots undergo this training.

Although I do not disagree with the results of that case, there is no way the holding can be reconciled with the Griggs disparate impact analysis. There was no showing of job relatedness. There was no empirical study as preferred in the Guidelines.

The Court did cite the Guidelines and pointed out that it had a provision--back at that time it was the EEOC Guidelines, but it is incorporated in the Uniform Guidelines -- that said if there is a high risk to society, a lesser demonstration of job relatedness is necessary. There is a balancing between the risk to society in terms of loss of life and property damage as opposed to how much validity evidence is needed. However, in Spurlock there wasn't any evidence of validity other than face validity. In other words, the Court accepted a face validity claim because it perceived the job to be difficult and dangerous to society.

The courts have been less than consistent about instances when high degrees of validity are required and when low degrees of validity are required. It is very easy to conclude an airline pilot has a

substantial responsibility to the public for the safety of air traffic passengers, etc., but when we examine truck drivers -- truck drivers who may be driving highly volatile or toxic chemicals through major metropolitan areas -- we are not willing to apply the same deference to the employer's wishes.

Just recently, a district court in Georgia struck down a one year experience requirement used by a trucking company to hire new truckers on the ground it disproportionately excluded women. The court used the disparate impact theory to strike down the one year experience requirement to be a truck driver and found there was no evidence the experience requirement was job related.

If the principle of face validity is going to be accepted, the one year experience requirement is equally no more face valid for the truck driver than the college degree requirement is for the airline pilot.

There are a number of other decisions which have come out of the courts without any genuine compliance with the Guidelines requirement of demonstration of job relatedness.

In the New York Transit Authority v. Beazer case, the Transit Authority refused to employ individuals who were on methadone programs of their prior drug addiction. The Court had no trouble finding the requirement was a legitimate job related requirement. It simply concluded the safety considerations of the transit authority were paramount.

Another interesting case is the Fifth Circuit decision in Smith v. Olin Corp. The employer proved the individual was not hired because he had a weak back. The individual claimed he had sickle-cell anemia and that was why he had a bad back and, therefore, alleged a disparate impact cause of action on the weak back theory.

The Fifth Circuit ultimately said there were some things which were so "obviously job related" that one didn't need to have a validity study to prove it. The Court concluded that having a strong back was obviously related to working in a chemical plant as a laborer. While the decision makes sense, unfortunately, the Court did not explain how to determine when this exception would apply.

The second issue of substantial concern is the cost of all the job related studies which have been generated because of the Griggs v. Duke Power rule. This issue is in part interrelated with the "professional standards" issue which is discussed later. When the cost to our society of some of the major validation studies and the litigation resulting therefrom is weighed, it is doubtful this approach makes sense.

Consider the New York State troopers case from several years ago. It was published in the press the validation study cost a million dollars and was paid for, at least in part, by federal tax dollars. Shortly after the test was implemented, a suit was filed by the federal governement--spending more of our tax dollars, seeking to enjoin the use of the test. The test was thrown out. So there was no real use of this million dollar test.

It could certainly not be established the state police, in that case, had not made a serious effort to validate a test--they spent a million dollars doing it. It was not that the test they formulated was terrible, it just didn't, according to the Court at least, quite meet the requirements of the Guidelines.

Observe the New York Correctional Department's Civil Service examinations. The Department has been litigating test validity for 15 years. During more than six years of litigation, the New York Correctional Department and the Civil Service Department lost on the Correctional Department Sergeant's exam. It took them only three years of litigation to lose the next case on the lieutenant's exam.

The New York Correctional Department and Civil Service Department then turned to the Captain's exam. They completed a validity study and proceeded to administer their test. Then they decided to be smart. To insure there was no adverse impact, so they did not have to face the validity challenge, they equated the means of the white and the black samples and used the scores from the adjusted means (of course before they could get their selections in they were nevertheless sued, this time by the white officers claiming reverse discrimination.) The district court enjoined their action and the appellate court remanded the case.

No matter what they do they can not seem to create a lawful exam. The essential result is the Correctional Department is run by court order because the only way the Department can promote someone is with court approval. Many Civil Service Departments all over the country are being run by court order. This does not make sense.

The third problem is the requirement that these validity studies demonstrate job relatedness in accordance with "professional standards." This requirement reminds me of my attempt to play the new video games. I have enough trouble trying to hit a dart board with the darts. "Professional standards" appear to move in many different directions at once, frequently changing galaxies, if not dimensions, and often they are more related to the particular spokesman who is "interpreting" the "professional standards," than to any independent observable set of rules. The mere comparison of the various drafts of the new APA standards is sufficient to show there is substantial disagreement as to what is "standard." Suffice it to say that the descriptions of the definition of what is "professional standards" have a very high standard deviation. Some of the drafts of the standards appear to require nothing less than the search for the Holy Grail.

This search for compliance with professional standards has also led to ridiculous results. Remember just a few years ago when there were a number of articles in the literature concerned with whether or not behaviorally anchored rating scales were better than other scales. Then suddenly, three systems appeared claiming they were superior to the BARS system. Study after study followed. There was never any consensus as to whether these scales were better or worse than any other scale. It seemed the entire focus of these studies came to no agreement, no consensus on "what was the professional standard."

Another example: I once had an I/O psychologist tell me an entire validity study was no good because the instructions on the criterion forms had been handed to each rater rather than being taught to the raters in a classroom setting. He did not have any problem with the content of the instructions, he simply said it had to be taught to the raters rather than merely allowing them to read it. That was his interpretation of "professional standard."

We need only look at the differential validity issue to see just how widely professional standards can swing. In the mid-60's when the first Guidelines were being drafted, the concept that tests worked differently for blacks, or didn't work for blacks and only worked for whites, was quite popular. This concept was espoused as fact by a number of the "leaders of the profession," and was thus incorporated in the Guidelines. After 10 years of research, the journals have now published numerous articles showing the theory was incorrect. During these 10 years, the theory was placed into law. It still exists in a form in the Guidelines. It still is included in a number of circuit court decisions. We have this legal requirement which has now been rejected by the weight of professional opinion. Will this change again 10 years from now?

If business tried to comply with all standards as they change, it could spend all of its time accomplishing that and none of its time operating its business. They would be good for job security for psychologists, but not very productive.

So, for all these reasons, the Griggs guidelines disparate impact rules are difficult to apply and really do not make sense for our society. Mostly they do not make sense because they do not fit the way people do business in this country.

I would like to see the courts and the regulators adopt a different standard altogether. In this standard of my dreams, adverse impact would be relevant, but it wouldn't be the trigger mechanism it is now. Under Griggs you either have adverse impact or you don't. And if you have it you must show job relatedness, and if you don't there is no requirement at all.

I envision the courts examining the business reason, the rationality of the business reason for use of a selection procedure when the procedure is attacked as being discriminatory. I would reintroduce the concept of intent into the analysis. Under my theory, the degree of disproportionate exclusion would be relevant but not controlling. The black who was not hired as a clerk typist because he couldn't stand on his head could win even in the absence of substantial disparate impact.

The demonstration of job relatedness would be required--but the more evidence of discrimination, the more substantial the evidence of job relatedness is required.

The question the court would focus on is one that frankly the employer should focus on and that is: Is there a rational legitimate business purpose for this selection procedure and is that why the employer is using it?

Once there is a rational demonstration of why it's important, the absence of intent to discriminate is shown.

The issue should not be whether a company "complied" with professional standards. Rather, the question should be, did the business do a reasonable study (not state of the art, not the best according to some post hoc review) and did it identify important company objectives demonstrating the need for the use of the test.

From this analysis one can determine whether the business reasonably adopted the selection procedure for business purposes rather than as a pretext for discrimination.

Businesses in this country should not be forced to mold their procedures to any one system. To the extent the courts are given freedom to address the particular employer's needs within a zone of reasonableness, I think that is the appropriate analysis.

This is better than making the artificial "calculation" of adverse impact determinative and then requiring the mystical search for compliance with "Professional Standards."

Once again, I'm not totally naive. I don't really believe the law is going to change in this respect. The disparate impact Griggs analysis is well embedded in our law, it is certainly embedded in the Guidelines.

On the other hand the law in this country is well known for moving in mysterious ways. It may well be that the equivalent of this proposed alternative can be achieved in another way, if I might just speculate for a minute.

Under the Griggs analysis, if there is adverse impact, the burden shifts to the employer to demonstrate job relatedness and then if the employer does that, the plaintiff has the opportunity to come back and prove

the use of the selection procedures was merely a pretext for discrimination, at least with respect to the use of paper and pencil tests. If the Schmidt and Hunter generalizability theory is proven in the courts in a series of cases and becomes generally recognized, then what you have is a situation where the law would start to recognize job relatedness on a relatively minor showing. So if the plaintiff proves adverse impact, the employer simply copies the Schmidt and Hunter articles, puts them into evidence and has established job relatedness. Certainly it's not quite that simple, but compared to what some employers have had to go through to demonstrate validity today in the courtroom, its relatively easy.

If that is accepted as sufficient proof of job relatedness, then the whole issue in the legal context is going to focus on the "pretext" issue. The pretext issue, when analyzed, boils down to whether the employer made a justifiable, rational, business decision. That is, was the use of the selection procedure important to its business objectives and, therefore, most likely not a "pretext" for discrimination.

So the analysis may, in fact, become a question of the rationality of business decisions. This analysis is superior to the technical requirements of complying with "professional standards", whatever they are, and is a standard which would better serve our society.

* * *

AUTOMATED TEST GENERATION (Paper and Demonstration)

Chair: Beverly G. Corkerin, Pinellas County, FL

Robert Maurer, New Jersey Department of Civil Service,
Trenton, New Jersey

I. Background

Production Related Statistics (How Much Do We Do)

The New Jersey Department of Civil Service, unlike most such organizations in the U.S., provides personnel services to county and municipal government agencies as well as state agencies. Our clients total 300 jurisdictions/agencies, and 200,000 employees. The Division of Examinations, which is responsible for the employment selection portion of the personnel program, has 150 employees and develops/administers approximately 1,500 separate assembled tests per year. Our current item bank houses about 75,000 items.

Current Operating Characteristics (How Do We Develop Tests)

Prior to developing a selection test for any title, an examiner researches the history of the title and determines the value of available job analysis information. If the information is still accurate and complete, the old job analysis will be reused. If not, a new one will be done. The job analysis method and information gathering technique used in any particular case are left to the professional discretion of the individual examiner.

When the job analysis is completed, test development work begins. There are a variety of item display vehicles in use, ranging from individual 4 x 8 cards to paper subtest forms to CPT word processor output. Regardless of the display vehicle, a great deal of examiner time is required to locate and cull the items, and a great deal of clerical time is required to transpose the material into camera ready copy.

Current Organizational Characteristics (What Do We Look Like)

On paper, the Division of Examinations is a typical government bureaucracy: Director -- Deputy Director -- Assistant Director -- Section Supervisors -- Team Leaders -- Workers. In reality, the structure is much flatter, allowing for a great degree of cross unit communication and formation of a large number of ad-hoc task forces. This is so for four basic reasons:

1. the work performed is extremely complicated, both technically and operationally

2. the workers are very well trained/knowledgeable in their fields

3. the workers prefer a decentralized work program which challenges their knowledge, discretion, etc.

4. without a rigid hierarchy, technical and procedural changes are easier to initiate; the demand to do more with less is causing us to change radically in short time spans -- Which leads us to --

Need for More Efficient Operating Characteristics (Why Did We Buy ATG)

The Division faces a large backlog in tests to be announced. In addition, there is an ever increasing clamor from appointing authorities and applicants to speed up the employment list production process. Finally, the current administration has cut our number of authorized positions and has not indicated that diminished service will be tolerated.

II. General Action Plan For Dealing With Our Situation (4 Primary Elements)

1) Generic Testing

The Division has moved away from the use of position specific tests, and has begun concentrating instead on identifying and measuring worker characteristics which are common across jobs. This allows us to make maximum use of each instrument developed, while not sacrificing validity.

2) Human Resources Planning

The Division has identified a rather crude human resources planning mechanism as an efficiency measure. That is, when a vacancy occurs in a particular title in a specific jurisdiction, we will announce a test for all jurisdictions using that title, regardless of whether or not vacancies exist there. In this way, we will develop a test only once and maximum use will be made of it. Another advantage of this approach is that it will minimize the chances for provisional appointments since lists will already exist.

3) Maximum Use of Unassembled Examinations

Specific criteria were established for determining when announcements could be processed via unassembled exams. By recognizing the legitimacy of this approach in a wider variety of circumstances, we have reached the point where a significant proportion of the employment lists issued by the Division results from unassembled exam procedures.

4) * More Efficient Test Production Methods *

The manual paper processing method I spoke about earlier contributes a great deal to our inability to deliver expeditious service. In addition, it comprises the more tedious elements of the production process; therefore, its cumulative negative effects are probably far greater than routine time estimates might reveal. At any rate, the remainder of this presentation will focus on our effort to reduce test production time by eliminating its more tedious elements.

III. Chronology of Automated Test Generation (ATG) System Installation

Assessment Systems, Inc. (ASI) and their ATG system were "discovered" by our Chief of Data Processing in mid-1983; shortly thereafter, the Department contracted for the system. A committee of examiners was then formed to review and evaluate ASI's functional specifications, and to make recommendations to upper management. Further meetings were held to negotiate modifications in the specifications; after agreements were reached, installation began (concurrent with the above activities was the development, by examiner committee, of item, job behavior, and worker characteristic taxonomies).

Installation consisted of the following major activities:

- programming
- getting up a model system comprised of one workload
- development of a user's manual
- orientation/training for examiners (the phase we are in now)
- implementation

IV. Major Managerial Concerns To Be Addressed By Any Organization Planning On Installing Such A System

Organizational characteristics must be considered in selecting and installing such a system. In our case the system was selected without consulting those who would use it. Given the organizational characteristics I outlined earlier in this presentation, this caused some very serious problems during the early phases of the installation process. To turn this around, examiner staff were heavily involved in the later phases of the installation process:

- a well respected journey level examiner was selected to function as the full-time project leader

- one of our Test Development Supervisors was selected to work directly with ASI on the development of the User's Manual.

- a regular newsletter, called the ASI Update, was instituted to keep all staff members informed of progress.

- committees of journey level examiners were established to plan and implement ancillary projects (job behavior, worker characteristic, item taxonomies - describe the reasons for creating the taxonomies)

V. Major System Features

Security

The system provides four levels of access: system manager, delete, modify, inquire. The "system manager" can perform all functions associated with ATG, inquire and change passwords, and set up user accounts and levels of security for personnel. Those at "delete" level can perform all functions of the system manager except maintain security. Those at "modify" level can inquire and modify records in ATG. This is the working level. Those at "inquiry" can view entries in the system.

Linkage

The system provides for the linkage of job title, job behavior (JB's), worker characteristics (WC's), and test items. This feature, combined with our taxonomies, minimizes duplication of effort in the generation of JB's and WC's, and, an important advantage, of course, is that the system linkage minimizes errors thereby contributing to the validity of our instruments.

Working Files

The working file allows each examiner to isolate his current work from the data base, thereby "protecting" it from "outside" interference. That is, an examiner can modify all items withdrawn to his work file in any way he sees fit without affecting those items in the data base or in any other examiner's work file. The idea of the working file is consistent with the Division's emphasis on decentralization.

Note Pad

Examiners can write notes to be associated with a particular test, and be assured that the notes will go to the test's history file. Such notes might describe problems, appeals, major successes, etc. associated with the test.

History File

A copy of the test as it was administered (regardless of subsequent changes made in items included in it) is retained in the history file. The history file can be reviewed on line; although the note pad can be modified, the test cannot.

User Friendly Orientation

The system provides very specific, easily read menu screens. There are also a number of simple messages displayed after certain function keys are depressed to minimize serious errors. For example, an examiner can break links by pressing an "unlink" function key. In such cases, the system responds with the message: "Do you want to break the current link?" The examiner must then respond positively before the link is broken. If "Yes" is selected, the message: "Link Broken" then appears.

VI. System Hardware

We now have a PRIME 9950 Supermini; we plan on adding a 9750 this summer. We now have 32 PRIME PT200 Terminals, and plan on adding to this number until each examiner has his/her own. It is important to note that this hardware is not dedicated to the ATG system, but rather services it along with an office automation system, and other ancillary systems.

VII. System Demonstration

As part of an overall effort to improve operating efficiency, the New Jersey Department of Civil Service contracted for a system which eliminated the more tedious, manual elements of the test production process. The system provides for access security linkages among job title, job behaviors, worker characteristics, and test items, working files for examiner "privacy", history files, and user friendly orientation.

Introduction: Throughout the demonstration, note the user friendliness of the system. Unless you are actually inputting items or job analysis data, there is very little typing required of the user. Instead, the work is accomplished simply by pressing appropriately labeled function keys. Note also that the system flashes instructions to the user in the lower left corner above the function key labels. It also confirms that certain actions requested by the user have been accomplished.

Today's demonstration will not be a full system demonstration because there isn't enough time for it. Rather, it will be a functional demonstration of how a typical examiner might use the system to develop a test. Therefore, only some of the screens will be demonstrated. If you want more information or a more complete individual demonstration, please see me or Steve Nettles before tht conference ends.

We will start the demonstration from the system's Main Menu, which is what the examiner sees when he initially signs onto the system. Note that the Main Menu contains the 4 basic elements which examiners work with on a daily basis:

1. Job Titles
2. Job Behaviors (You may call them task elements or activity statements in your jurisdiction).
3. Worker Characteristics (You may call these knowledge/skills/abilities in your jurisdictions).
4. Individual items and complete examinations.

Keep in mind that we can consider the system to have two major subsystems, and so we will orient the demonstration in that direction. Again, we will approach the demonstration from the point of view of an examiner who has been assigned to develop a test for a particular title -- and we have selected the title "Filing Clerk" for today's demonstration.

I. **The system can be used as an automatic test generator -- we will use the Title Screen to demonstrate how one of our examiners might complete his assignment by researching the title's history (includes most recent job analysis information and all previously administered examinations). From the title screen, the examiner would:**

   A. **press the key corresponding to "Inquire" because he wants to research the history of the Filing Clerk title.**

   B. **reference this title by inserting its official title code. The title code is the title's official, unique, numeric identifier. The PAS Group Code, which you see off to the top right, is a taxonomy code which an examiner might use to identify titles which are similar to the one he is interested in. If there is no history available for the Filing Clerk title, there may be something available for a very similar title which he may use.**

A. select the Job Behavior screen from the Main Menu.

1. press "Inquire" to indicate that he wants to search for particular types of JB's.

2. insert the Taxonomy Code(s) which the examiner's research.

3. for each JB having the Tax Code entered, the examiner may:

a. link it to the title by pressing "Link JB" (demonstrate only once)

b. go to the next record (i.e. bypass or not use that JB). We will go through 3 or 4 of these to demonstrate the process.

B. select the WC screen from the Main Menu.

1. press "Inquire" to indicate that he wants to search for particular types of WC's.

2. insert the Taxonomy Code(s) which the examiner's research.

3. for each WC having the Tax Code entered, the examiner may:

a. link it to the title by pressing "Link WC" (Do not demonstrate -- essentially the same as the JB link process).

b. go to the next record (i.e. bypass or not use that WC). We will go through 3 or 4 of these to demonstrate the process.

Note: JB's and WC's cannot be moved to workfiles, and so any modifications or deletions of them take place in the data base. Therefore, these functions regarding JB's and WC's are reserved for supervisor's.

C. select the Exam Generation screen from the Main Menu

1. enter a new booklet ID and administration date for the test to be developed.

2. enter Tax and Ability Codes of items he will want to work with (Note that Tax Codes of items correspond to those of WC's so once the examiner has identified relevant WC's, he knows where items may be drawn from). He will also enter the number of "New" and "Used" items he wants to use in the test. Note that the number available in each of these 2 categories is provided by the system.

C. press "Other Options".

D. press "Booklet History" to see if there were any tests administered for this title in the past. And the system shows us that there was one.

E. The examiner will press "Exam Review." At this point, the examiner may review the items that were included in the prior test. Note the labels running across the top of the screen:

- the ID code is the item's unique identifier.
- the Tax and Ability codes represent categories which we have assigned the items to.
- the Reference code indicates the item's source.
- the Author code indicates who wrote the item.
- the Status code indicates whether the item is still unapproved or has been moved to an approved status.
- the Key, Date Created, and Last Changed are self explanatory.

By pressing the "Worker Characteristic" function key, the examiner may see the WC's which the item was linked with in the past. Further, by pressing the "Job Behavior" function key, he may see the JB's it was linked with.

We will cycle through a few items and their corresponding WC's and JB's just to determine this process.

F. If the test in history looks like it might meet the examiner's need, press "Exit Select".

G. press "Clone HistFile" to copy the test to a workfile -- remember, the workfile is the individual examiner's "private" property. He moves the test to a workfile because he cannot modify history.

H. press "Review WorkFile" after history is cloned.

I. cull the items, modifying as necessary to meet the immediate need. By pressing "Other Options" at this point, the examiner may also add or delete items (do not demonstrate); may ask to review any notes that were associated with this test in the past (demonstrate); or may ask for the item's statistical history (demonstrate).

J. The system can be used as an information bank to be used by the examiner when no specific or related test history exists for a title which he has been assigned to work on. (Tie into title codes and group codes) In this sub-system, the examiner might proceed as follows:

After the examiner specifies these parameters, the test is built automatically by the system in accordance with the parameters. Note that "New" items are selected at random from that category; "Used" items are selected on the basis of prior use: least used, first selected. Note also that the system only selects items having an approved status.

3. press "Create WorkFile".

4. press "Review WorkFile" -- from here, the examiner may:

   a. go to the next item if he likes this one. Let's go through 3 items to demonstrate the process.

   b. modify an item (without disturbing the item bank itself -- remember that no one else has access to the workfile). We will demonstrate modifying the 3rd item. Note the system's response to the command in the lower left corner.

   c. select other options: (We will discribe, but not demonstrate the functions):

      -- delete the item
      -- add an item
      -- write a note to the exam note screen
      -- get the item's statistical history if it has been used before.
      -- clone a new or modified item, thereby placing it in the data base as well as in the workfile.

5. press "Print/Typeset" for a clean and keyed copy of the test. We will distribute copies of what printouts might look like.

* * *

* * *

PUTTING VALIDITY GENERALIZATION AND TRANSPORTABILITY TO OPTIMAL USE (Symposium)

Chair: Jeffrey P. Feuquay, Oklahoma Office of Personnel Management

Application of Validity Generalization Within the United States Employment Service

Jerry W. Pickett, Employment Security Commisison of North Carolina, Raleigh, North Carolina

Labor exchange operations of the United States Employment Service (USES) and affiliated local Job Service offices function to bring job seekers together with employers seeking workers. To screen applicants for job placement selection, and to counsel for career alternatives are difficult processes requiring considerable technical skill and objectivity, and interviewing alone is insufficient to satisfactorily achieve those goals.

In the past, some problems with USES tests have been:

- Coverage. After many years of research, only about 450 jobs have been covered of the 12,099 in DOT. Jobs are being created faster than we can research them.
- Selectivity. We can identify those who will probably fail, but cannot identify the most capable using current techniques.
- Fairness. Some individuals feel that all tests are unfair to minority groups.

Recent technical advances have permitted the use of our massive validation data base to address some of the basic problems. These findings are based on huge amounts of data, using technically rigorous, state-of-the-art analytic procedures.

The findings are extremely stable and dependable.

- The General Aptitude Test Battery (GATB) is valid for all jobs in the Dictionary of Occupational Titles. Jobs can be grouped into five broad job families. Validities within these JOBFAMs are very similar; validities differ between JOBFAMs.

- The GATB can be used to identify the most capable rather than only the minimally compentent.

- The GATB is fair to minorities; a given test score is associated with virtually the same level of job performance regardless of the race of the person. Also, the tests can be used in such a way as to eliminate adverse impact -- the race of referrals mirrors that of the applicant population.

- The GATB can be used in a such simplier and effective way.

Validity Generalization (VG) is a new method of using existing tests to improve the accuracy of referrals. By referring the applicants who will be most productive, this new method is expected to increase employer acceptance, ES effectiveness, and the productivity of employers' work force and thus increase the productivity of the economy as a whole.

Limiting factor in ES productivity/performance is the number and quality of job orders.

Vicious circle: We don't get good job orders (or any job orders from some employers) because the employers perceive that we send them incompetent applicants. The best applicants may not apply to us because they believe that we have only low quality jobs.

Solution: Refer the best applicants to all job orders in order to break the cycle. We have capable people — the problem is how to identify them, i.e., quality screening.

Operational Implications

Optimal selection for referral requires the following steps to be taken:

- Use the GATB to test virtually all applicants.

- Use the GATB as the primary selection factor for referral to virtually all jobs.

- Consider all applicants for all job orders, i.e., search all files.

- Refer the most able applicants to each job order.

Basic Approaches

1. Full implementation: When this approach is adopted, the GATB is administered to the majority of new and renewal applicants. Test selection is heavily promoted to private and public employers. Test results are considered when making almost all referrals. Percentile score reports are sent to employers who

request them, and referrals for most other job orders are based on top-down selection using the percentile scores.

2. Partial Implementation: For this approach, test selection is promoted to private and public employers, and the applicants for job orders requesting test selection are tested. Percentile score reports are sent to employers who request them.

Operational Experience

Because economic and technical justification for VG are compelling, a pilot test of VG was initiated in North Carolina to develop operational procedures and to evaluate impact on performance. The evaluation period was FY '82, a time of declining economic conditions, stagnant hiring patterns and Employment Service staff cuts. Even so, the results were uniformly positive.

Evaluation of the project consisted of:

(1) Comparison of ESARs data on production, penetration, testing, placements, etc. for VG offices against statewide totals and matched offices. Matched offices were selected on size of office and UI rates.

(2) Employer survey designed to measure impact of VG on hiring practices.

(3) A case study of staffing the Philip Morris Cabarrus facility.

- ESARS Data

Selected results of comparisons of VG local offices with other local offices and to a matched group of offices for FY 1982 are as follows:

	VG	Other	Matched
Number of Placements	+ 4%	-22%	-20%
Penetration Rate	+23%	-13%	+ 1%
Productivity	+18%	- 6%	- 6%
Referrals Made	+ 5%	-17%	-22%

We must remember that the period of evaluation was during a period of economic decline in N.C. as well as most areas of the country. A 4% increase in placements may be small, but in view of the large percentage decreases for the comparison groups, the 4% increase can be considered significant.

- Employer Survey

The Employer Survey consisted of 195 employers who were familiar with the VG testing in the Raleigh-Durham job bank area. A few highlights of the survey are:

(1) 50% of employers thought VG had made Job Service (JS) more useful in their hiring process;

(2) 36% employers had hired VG referrals for different types of jobs since VG;

(3) 50% of employers hired a higher percentage of referrals under VG;

(4) 61% of employers who used VG said it had saved them money in training and other personnel costs;

(5) 79% of employers who had not had the opportunity to use VG said they thought VG could save money in training and other personnel costs;

(6) 55% of employers said VG had encouraged their willingness to enter a sole source hiring agreement with JS.

- Case Study: Philip Morris

The study by Philip Morris simply confirms VG factual premises. Philip Morris compared applicants screened by VG against transfer applicants not test selected, their other successful operating plants, and National averages. They were measured for effectiveness on training success, disciplinary action, safety, quality and production. The company concluded that, "... out of the 14 comparisons for which data was available the GATB screened new plant employees exceeded the comparison groups in 13 ..." Of the (13) comparisons the average "improvement" margin was a whopping 41 percent".

The evaluation made by the employer was as follows:

> Overall, the new hires have more than exceeded expectations, and have created a workforce which can be characterized as faster learning, more disciplined, safer, more quality conscious and more productive. We would again like to thank the NCESC for their valuable assistance in helping us to select such a higher caliber of employees.

Dr. Michael McKinney, the outside consultant who directed the North Carolina evaluation study, estimated the total value of increased productivity due to VG in the pilot area to be some $21 million and the cost benefit ratio of testing costs to increased productivity to be $1 to $225.

Because of the positive operational experiences, favorable hard data and enthusiastic employer response, North Carolina was the first state to expand the VG testing system statewide to all Job Service offices.

In order to verify the North Carolina results and to develop operational procedures for other settings, several other pilot projects were initiated at State request. Roanoke, Virginia, the most advanced at this point, has built upon the North Carolina experience. Observing that North Carolina had not been able to do the optimal amount of testing, Virginia has introduced operational efficiencies such as group applications, interview scheduling, mass testing, microprocessor-assisted file search and test scoring, and discouraging repeat visits. Virginia has been able to test over 75 percent of their applicants and are using VG in almost all referrals. During the March-December 1984 period, placements increased 18% (versus 3% Statewide) the fill rate was 60% (versus 50% Statewide) and 69% of referrals were within three days of receipt of the job order (60% Statewide).

In addition to North Carolina and Virginia, a number of other States have requested and received authorization to conduct VG pilot projects.

VG is <u>feasible</u> at current staffing levels given:

- Scanning
- Modified applicant flow including less interviewing time and other time-savers such as self-application and, ideally, automation.
- Full commitment by top management

Most of the VG projects are not yet in full operation; their full impact has yet to be felt. Even so, employer interest is apparent and seems to be increasing. This is particularly notable in the automobile industry where Chrysler has hired a considerable number of workers using VG, and General Motors has requested and used VG in several States. Three other large companies (Philip Morris, Hercules and Utah Power and Light) are known to have made multi-State requests. There may well have been others. It is certain there will be more.

Note: A very comprehensive reference of technical reports and bibliography were supplied by the author.

* * *

VALIDITY GENERALIZATION SUMMARY

John Hawk and Jerry Pickett, United States Employment Service;
and Richard C. Gilliland

Background

The State Employment Security agencies have used ability tests since the mid 1930's to select applicants for referral to jobs and as an aid in rational career decision making. Throughout this time the Department of Labor (DOL) has conducted a vigorous and productive research program which has created and developed such highly respected instruments as the General Aptitude Test Battery (GATB). This battery of test measures a very wide range of cognitive, perceptual and psychomotor abilities and has been found useful for many diverse jobs in a wide variety of settings, including foreign. One function of the research program has been to determine the validity of the GATB, i.e., the relationship between GATB test scores and job performance or productivity.

Using the best research methods available, a great deal of information was collected. However, because each research study related only to the specific occupation studied, only some 400 of the more populous jobs were covered of the more than 12,000 in the economy. Also, the interpretation of the GATB provided only three categories: high, medium and low.

Recent advances in analysis, known collectively as meta-analysis or Validity Generalization (VG) have radically altered this situation; the basic testing needs of the Employment Service can now be met. Using the state-of-the-art VG techniques on the massive DOL data base have resulted in a number of very important research findings with pervasive implications for the public employment service. The prevailing consensus within the field of psychometrics went through the following three steps.

Step I -- A test which could predict success in a specific job in one location would not necessarily predict success for the same job in a different location. This was the thinking in the 1930's and 40's.

Step II-- Test validity can be generalized for a specific job regardless of location. The use of Specific Aptitude Test Batteries (SATBs) from 1947 to date is based on this formulation.

Step III- If a test is a valid predictor for some jobs randomly selected from a larger cluster of jobs, and there is little variation in the validities, then the test is valid for all jobs in the larger cluster. This larger cluster of jobs in which worker functions are the same and aptitude requirements are similar are called job families. Research shows that all, e.g., assembler

jobs in the economy would fall into the same family. The concept of generalizing the validity of a test based on relatively few studies to a larger number of equally complex jobs is called Validity Generalization (VG).

The research which allowed us to move from Step II to Step III was conducted by one of the leading authorities in the field, Dr. John Hunter, in conjunction with the staff of the Southern Test Development Field Center. The research is based on the cumulative GATB validity research data amassed over many years which was analyzed using state-of-the-art statistical techniques developed by Dr. Hunter and his colleague Dr. Frank Schmidt of the Federal Office of Personnel Management. Research on this cumulative data base also reveals that gender, age, ethnic group, and geographical areas of the country have no discernable impact on validity. Thus, validity is generalizable across types of applicants as well as jobs within job families.

Composite Scores

Based on the work of Dr. Hunter, nine aptitude scores of the GATB were combined and reduced to three composite factors. They are: a Cognitive Factor comprised of verbal and numerical skills; Perceptual Factor comprised of spacial, form perception and clerical perception skills; and a Psychomotor Factor comprised of motor coordination, finger dexterity, and manual dexterity. This was done because it is the general cognitive ability, the general perceptual ability, and the general psycholotor ability that best predict job success. The generalized abilities are stronger predictors of job success than are all of the aptitudes working independently

Job Family Concept

In order to determine what job groupings would best meet the goals of validity generalization and maximization, five methods of grouping jobs were examined by Dr. Hunter. All five worked in the sense that they produce greatly expanded occupational coverage and very useful levels of prediction. The method of job grouping judged to be best consisted of five job families covering all jobs in the DOT. This grouping (Job Families) is based on the complexity level of the Data and Things function of the DOT code and was originally suggested by the staff of the Southern Test Development Field Center.

The meaning of these findings is that the GATB can be used to predict success for any job in the DOT with very high confidence--much higher than with the current system of Specific Aptitude Test Batteries.

Relationship Between the DOT Code and the Five Job Families

The middle three digits of the nine digit DOT occupational code are the worker functions ratings of the tasks performed in the occupation. Every job requires a worker to function to some degree in relation to data, people and things. A separate digit expresses the worker's relationship to each of these three groups: data (4th digit), people (5th digit), and things (6th digit). Worker Functions involving more complex responsibility (such as synthesizing date, negotiating with clients, or doing precision mechanical work) and judgement are assigned lower numbers while functions which are less complicated (such as copying data, serving people, and handling objects) have higher numbers.

The Table below shows how the 12,000 jobs in the DOT are grouped into the five Job Families using the Data and Things code of the DOT. Analysis showed that the people function (fifth digit) did not add any additional validity to the Job Families.

JOB FAMILY CLASSIFICATION SYSTEM

Job Family	Description	DOT-Fourth or Six Digit	No. of DOT Jobs
JF I	Set-up; Supervisory	Things 0	398
JF II	Feeding-Offbearing	Things 6	400
JF III	Professional; Supervisory	Data 0,1	2103
JF IV	Skilled Trades; Clerical	Data 2,3,4	3145
JF V	Semi-Skilled Trades	Data 5,6	6053

Composite Contribution to Job Families

The analysis also showed the importance of each composite score within a Job Family. For example, 59% of the Job Family I score is based on GVN (Cognitive), 30% is based on SPQ (Perceptual), and 11% is based on KFM (Psychomotive). The results by GVN increases. Using calculated weights from the composite contribution analysis, a percentile score is computed for each individual for each of the five Job Families. In order to eliminate adverse impact, within-group percentiles are computed for Blacks, Hispanics and non-minorities. These percentile scores are directly related to job performance. Earlier research had shown that for every point increase in test score, there is a corresponding increase in average job performance. Thus the percentile scores provide much finer meaningful than the High, Medium and Low catagories produced by SATBs.

Dollar Value

The projected economic impact of Validity Generalization is far higher than the public and the employer community imagines. Dr. Hunter found that the value of USES test, as currently used, have a dollar value to our employers and the economy as a whole of 1.73

billion dollars per year. This value, as large as it is, can be increased dramatically. If the number of test-selected placements remains constant, but the tests were used in a more optimal way by using composite aptitude scores and raising the cutting score so that the selection ratio reflects our applicant to placement ratio, the value would increase to 7.94 billion dollars per year. If all placements, made by the Job Service system were test selected, the value to the economy would be 79.36 billion dollars per year!

Benefits of Validity Generalization

Several other benefits are produced by Validity Generalization. Benefits for applicants include: proper job choices eliminating lost time in jobs for which they are not suited; more information to help counselees including the probability of success in specific jobs, as Validity Generalization is expanded to counseling; increased job satisfaction through increased probability of success; higher quality jobs provided through employer willingness to place a wider variety of job orders; and, no adverse impact. Benefits to employers include: valid tests available to cover most jobs in the job market; personnel administration cost is reduced; increased quality of workers; reduced turnover, supervision time, material waste, absenteeism, and morale problems; no adverse impact; and, signigicant increase in worker productivity in terms of quality and quantity.

There are also benefits for job service such as: increased job orders and placements; better referral/placement ratio; increased applicant renewals; better paying job orders with higher skill levels; additional information to help counselees will be available as Validity Generalization is expanded to counseling; and, more qualified applicants.

Alternatives to Tests

Analyses similar to the ones performed on the GATB validity base were recently reported by Dr. Hunter on the cumulative data base for other commonly used predictors. As the value of increased productivity due to good selection is directly related to the magnitude of the validity coefficient, the value of the GATB is roughly three times that of experience and four times that of the interview.

Operational Implications

Optimal selection for referral requires the following steps be taken: use of GATB to test virtually all applicants as the primary selection factor for referral to virtually all jobs; consider all applicants for all job orders, i.e., search the files; and, refer the most able applicants to each job order.

Operational Experience

Because economic and technical justification for VG are compelling, a pilot test of VG was initiated in North Carolina to develop operational procedures and to evaluate impact on performance. The evaluation period was FY '82, a time of declining economic conditions, stagnant hiring pattern and Employment Service staff cuts. The results were uniformly positive and indicated that employers are very favorably disposed toward the VG concept and that they seem to be changing their hiring practices to hire a larger percentage of Employment Service applicants. Also, many employers reported hiring ES applicants for different kinds of jobs than previously.

In order to verify the North Carolina results and to develop operational procedures for other settings, several pilot projects were initiated at State request. Roanoke, Virginia, the most advanced at this point, has built upon the North Carolina experience. Observing that North Carolina had not been able to do the optimal amount of testing, Virginia has introduced operational efficiencies such as group applications, interview scheduling, mass testing, microprocessor-assisted file search and test scoring, and discouraging repeat visits. Virginia has been able to test over 75 percent of their applicants and are using VG in almost all referrals. During the March-December 1984 period, placements increased 18% (versus 3% Statewide) the fill rate was 60% (versus the 50% Statewide) and 69% of referrals were within three days of receipt of the job order (60% Statewide). In addition to North Carolina and Virginia, a number of other States have requested and received authorization to conduct VG pilot projects.

Two patterns of VG use are emerging in the pilot projects. The first is the "full implementation" approach, exemplified by the Roanoke, VA project, in which procedures are modified to achieve maximum efficiency and are tailored around VG. The second is "partial implementation," or SATB replacement which is a demand-driven add-on to local office operations. Problems with the partial implementation approach include the lack of staff to do testing, the decision on which applicants should be tested, and the inherent inefficiency of considering tested applicants for only one or a few job openings. There seems to be a tendency for partial implementation projects to convert to full implementation. This is likely to produce operational changes similar to those introduced in the full implementation models.

* * *

MULTIPURPOSE JOB ANALYSIS (Symposium)

Chair: Marianne Bays, Prudential Insurance Company

Multipurpose Job Analysis Works, But

Gary B. Brumback, U.S. Department of Health and Human Services, Washington, D.C.

I will be talking to you today about the Department's multipurpose job analysis (MPJA for short) project, I am dedicating my talk to the memory of our late colleague and friend, Steve Bemis. More than a friend was taken from us by his abrupt death a few months age. We also lost an authoritative and active force in our field. He was an expert on job analysis and was to have been our discussant for this symposium. The last time I saw Steve was at the January luncheon meeting of the Personnel Testing Council of Metropolitan Washington. He was the guest speaker and talked about recent advances in job analysis[2]. In his concluding remarks, he turned briefly to the subject of MPJA and expressed some doubt over whether it would ever gain widespread use.

Steve had cause to wonder. MPJA does not dot any map of the organizational world.[3] Why that is so is probably due to the practical arguments against MPJA. I mentioned some of them in a progress report I gave at an early stage of our MPJA project, the 1983 annual meeting of this conference.[4] Facing such arguments plus what has seemed like daily difficulties in moving the project along has made it the most frustrating project of my long career. While not wishing to discourage you from considering MPJA, I must be honest and say I sometimes wonder if MPJA is worth the trouble, hence the hedge at the end of my talk's title. But I speak of my own personal frustrations and opinions. You must judge MPJA for yourself, of course, and I hope my talk will help you do that.

Here is what I plan to cover today: first, a quick recap of an overview I gave of the project start in my 1983 talk; second, a description of our job analysis methodology, including what we did not try, what we tried that did not work, and what did work; and third, a highlighting of the products we are producing and their application. Sprinkled throughout my talk will be little lessons we have learned along the way, should knowing them in advance be helpful to you. Finally, I am going to leave you with some heresy at the end.

Overview

From Idea to Project

Given the rationale against doing MPJA in the first place, you ask how it happened that we were able to initiate the project? The

answer is that a senior-level personnel administrator who believed more in the potential than in the futility of MPJA asked me to initiate it.[5] Lesson one: If a debatable idea belongs to a superior, the idea will likely go farther.

The objectives we set for the project were a) to develop an integrated methodology that would be truly multipurpose, including capabilities for meeting both employee selection and position classification needs, b) to analyze positions types within certain occupations in the Department using the methodology, and c) to develop generic personnel procedures for each of the position types based on the job analysis data.

We selected occupations which represent a variety of professional and clerical jobs and are populous and problematical enough in the Department to maximize benefits assuming the project would work.[6] The occupations are:

- Administrative Officer
- Clerk-Typist
- Computer Specialist
- Management Analyst
- Nurse
- Personnel Clerk/Assistant
- Program Analyst
- Secretary
- Social Insurance Administrator
- Social Science Analyst

There are many more occupations in the Department, but they are either not very populous, not very problematical or are concentrated in only one or a few of our component agencies.[7] We wanted dispersed occupations because the project is under the auspices of the Department, not any one of its agencies. The ramifications of a centralized project in a decentralized organization go beyond picking occupations, as you will notice throughout the rest of my talk.

Mobilizing Volunteers

Running the project is like running a church. Our agencies are my parishioners, not my subordinates. Lesson Three: It is much, much harder to persuade people to do things for you voluntarily.

There are two groups of volunteers. One is the project team made up of some 20 personnelists around the Department. They contribute time out of the goodness of their heart and only after making sure their regular duties don't get sacrificed. On behalf of the project this group tries to persuade the second group, the supervisors and employees associated with the jobs we are analyzing, to participate voluntarily in the job analysis.

The Contractor

The project team with my direction and help developed and pilot tested the original methodology.[8] We then turned it over to a contractor to use (and modify as necessary) in analyzing position types within the selected occupations.[9] The contractor is also developing the generic personnel procedures. Lesson Four: If you cannot mobilize the necessary resources in your organization, hire a good contractor.

Job Analysis Methodology

What We Did Not Try

Being an eclectic by nature and given a multipurpose need, there was very little methodology we did not try. We initially considered the Position Analysis Questionnaire (PAQ) developed by McCormick and his colleagues at Purdue University.[10] We did not try the PAQ, however, because it did not seem to be relevant enough (or at least fact valid enough) for our jobs and the Federal government's position classification system. A derivative of the PAQ, the Professional and Managerial Position Questionnaire (PMPQ) developed by Mitchell for his dissertation under McCormick, seemed more relevant.[11] I wanted to try the PMPQ, but most team members were skeptical about its applicability.

We did not try on-site observations of workers or participant observations. Observations can be useful, but we ruled them out as not being cost effective. We did not try several other approaches we spotted in our literature review. I will not bother to single them out here. What we did not find in the literature or in our calls around the country were any multipurpose methods to borrow, saving us the chore of developing our own. Oh, we found some methods, but one purpose or another would be neglected, and usually it was the position classification function.

What We Initially Tried

For the first part of our methodology, I decided to try CODAP (Comprehensive Occupational Data Analysis Program).[12] I imagine you are all familiar with it. My primary interest in CODAP was its ability to identify types of positions by clustering them into homogeneous groups and to array tasks statistically in a variety of formats. We developed a task inventory for one of the occupations and used Air Force version of CODAP to analyze the employees' responses to the inventory.[13]

The second part of our original methodology was the basic part. It involved a job analyst and a panel of subject matter experts (SMEs) going through a three-day process. We had developed it by merging

a selection-oriented job analysis panel method with additional steps needed to get data for our other purposes. The core method had previously been adapted by one of the project's team members from a method used by the U.S. Office of Personnel Management (OPM).[14]

Most of the three-day process involved procedures you probably have used in one form or another yourself. The FES questionnaire, however, was an innovation I want to single out for a moment. FES stands for the "factor evaluation system," the U.S. Government's official method for classifying its civil service positions. We developed the questionnaire by abbreviating OPM's primary benchmark descriptions of the factors and their subdivisions.[15] The job analyst and panel used the questionnaire in a group discussion to arrive at benchmark descriptions for the position type being analyzed. Our hope was that the questionnaire and group consensus would make the determination of factor levels more objective than the classifier's desk audit.

What We Learned (Lessons Five Through Ten)

The task inventory took entirely too long to develop, administer and to get the data ready for computer analysis. We ran into one delay when an advisory group of SMEs told us that no one would respond to an inventory with over 400 task items. So we took more time to edit the inventory, reducing it in length by almost two-thirds. Even then, we had to extend the deadline several times before we had accumulated enough usable returns. The experience caused me to abandon this approach for the remaining occupations and to rely on quicker ways to identify position types. But it is conceivable that we may someday return to task inventories. They would make an automated, task-based personnel system possible. I am not able at this time, however, to make a convincing case for the need for it.

My decision not to use task inventories again automatically ruled out further use of CODAP. I remain impressed by its many analytical capabilities, although it could not help us with the position classification function. The CODAP algorithm which arrays tasks by the different grade levels of the positions represented in the sample showed virtually no differentiation among the tasks. Of course, the general level of the task descriptions was partly responsible for the lack of differentiation. Yet tasks alone, no matter how specifically written, are insufficient to determine grade levels. Nine factors in the FES are required to determine a position's grade, and tasks contribute only part of the information needed. Obviously, CODAP cannot be blamed for what task data cannot do. Had our FES questionnaire items been ready when we put together the task inventory, the FES items could have been put into capability to analyze data from that section in combination with data from the task checklist section, we probably would have gotten some grade differentiating results out of the analysis.

We experienced several difficulties with the basic part of our methodology. First, the project team had a very hard time recruiting volunteer SMEs who could or would leave their work to sit on a panel of three, usually consecutive days. Second, many times we would find out during the first day that some of the panelists were the wrong SMEs. They were not doing or supervising the work represented in the position type and grade level we were supposed to be analyzing. We even had to abort entire panels and start over because most or all of the panelists were the wrong ones. The recruiters did not know beforehand which panelists were in misclassified, overgraded jobs. In fairness to the recruiters, the many shortcomings of the Federal government's classification system cause misclassification and overgrading in my opinion.[16] Third, the FES questionnaire caused lengthy debates over which benchmarks were the most appropriate ones. Fourth, cramming the multiple purposes into three days left too little time for any one purpose. And while most panelists said they enjoyed the experience, they and the analyst were usually tired by the end of the third day.

What We Are Doing Now

Our current process is broken up into three phases: 1) a background study of the occupation, 2) individual interviews and 3) SME panels.

The background study gives us a broad understanding of the occupation, including any personnel management issues associated with its use in the Department. The job analyst reviews a sample of PDs, position evaluation statements (official documents which explain the occupational designation of the position and its grade level), any classification appeal decisions, any earlier studies that might have been done on the occupation, and so on. The job analyst also sits down with project team members and other representatives of the personnel offices in the Department to make a preliminary identification of the position types, including grade levels for which generic personnel procedures would be the most useful. Sometimes, occupational representatives also are asked to advise us on the position types to be targeted. For example, we relied on an advisory group of nursing leaders in the Department to confirm our preliminary determination that 14 position types in the nurse series (e.g., ambulatory care nurse, nurse midwife, occupational health nurse, operating room nurse, and so on) should be studied.

The job analyst next interviews a sample of about 70 employees who supposedly occupy positions which represent the position types identified in the first phase. Employees in positions outside the types are sometimes interviewed also in case there should be noteworthy types we might otherwise miss. The interviews last about two hours, follow a protocol tailored to each occupation. The interviews constitute a miniature desk audit oriented toward the classification purpose with as much FES-related information being obtained as possible. Additionally, the interviews allow the analyst to screen out employees who would be the wrong SMEs for the panels in the third phase.

By taking the classification function out of the panel process, we have been able to reduce the time required for it down to about one and one-half days. Each panel has about four employees and four supervisors (but they do not supervise the employees who are on the panel). The panel's deliberations are aimed primarily at the employee selection and performance management functions.

Generic Personnel Procedures

The whole point of our MPJA project is to use the job analysis data to produce generic, or model, personnel procedures and to make them available to personnel offices and their clients (supervisors and employees). A generic procedure reflects the job requirements shared in common by positions which belong to a given position type. We currently are producing four kinds of generic procedures; position descriptions, position evaluating statements, crediting plans and performance plans.

The procedures are intended to save user's time. Since the job analysis and development work has already been done for the common job requirements, all a user has to do is determine if a specific position is represented by the position type and, if so, use the generic products. Of course, if the specific position also has some unique requirements that are important enough to take into account, the user would have to add to the generic product before using it.

The procedures are also intended to reduce classification disagreements and other unwanted differences in the way similar positions are treated. However, the Department does not force its agencies to use the procedures. Again, it is a voluntary matter.

Conclusion

Our project seems to be gradually gaining respectability as more people begin to see and use the products. So maybe the project has been worth the trouble after all. For the heck of it, though, I am going to close by playing devil's advocate. What I am about to say I would never have believed saying several years ago. My opinions seem to be changing for the better or the worse.

I used to urge job analysis, and a lot of it. I used to lament whenever management did not seem to appreciate or understand my urgings. Now, I wonder if I have been overselling job analysis. Here are three reasons why I wonder.

One. I am changing my mind because of the work Jack Hunter and his colleagues have been doing on validity generalization.[17] They have found what most managers have probably known all along; namely, there is no substitute for general intelligence in enabling a person to be successful on the job. Any one of numerous mental ability tests apparently would be the best predictor, and maybe the only one

necessary. A comprehensive, rigorous job analysis would be a waste of time. Simply find out if there is a match between the job to be filled and the jobs for which validity generalization data exist (and data exist for at least 75% of the jobs in the U.S. economy).

Two. The U.S. Government's position classification system needs to be radically changed. I run into very few people who have a good word for the present system. We need to stop making unmakable, fine-grained distinctions among levels of responsibility by either reducing the number of levels or by switching from a rank-in-position classification to a more performance-based, rank-in-person process for determining compensation.[18] Any revision of the classification system is bound to place less stress on the job analysis.

Three. Managing performance really does not require much of a job analysis. It can usually be very quick and informal. I think most managers are smart enough to understand and ensure that they do not hold subordinates accountable for work outside their official responsibilities. The only time serious job analysis needs to be done is when an organization wants to have generic performance standards for a class of jobs. And organizations should think twice before going ahead because generic standards are not without their disadvantages.

Well, I have spoken my mind. If I have stimulated you into reacting, isn't that better than to have put you to sleep?

FOOTNOTES AND REFERENCES

1. Paper presented at the annual meeting of IPMAAC, New Orleans, June 1985. The opinions expressed in this paper are the personal views of the author. No endorsement of them by the Department is implied or to be inferred.

2. Bemis, S. Recent advances in job analysis. Paper presented at the January 1985, meeting of the Personnel Testing Council of Metropolitan Washington.

3. Marianne Bays, our chairperson, tells me that speakers at last year's conference "were still calling unsuccessfully on their audience for examples of use of multi-purpose job analysis practices."

4. Brumback, G. and Palomino, P. Toward multi-purpose job analysis in a large public agency: Rationale, design and progress. Paper presented at the annual meeting of IPMAAC, Washington, D.C., May 1983.

5. The administrator and believer is Donna Beecher. You may know her or recognize her name. She is very active in IPMA, expecially its Federal Section of which she is a past President.

6. A problematical occupation is one which causes multiple headaches like classification disagreements, vague selection criteria and so on.

7. One of the occupations is unique to just one of our agencies, but it is a large occupation and the agency is our largest agency.

8. I wish to acknowledge the developmental work of these people in alphabetical order: Marcia Goldblatt, Janine Hornicek, John Nolan, Priscilla Palomino, William Scott, and Mike Turner.

9. The contractor is HGL Associates, Inc., of Arlington, Virginia. I am very appreciative of the quality of their work and their forebearance in coping with the many trials and tribulations of this project. Helen Liebman, Barbara Brock, Kathi Menda and Joe Cavallaro are the contractor team I would like to acknowledge here by name.

10. McCormick, E.J., Jeanneret, P.R., and Mecham, R.C. Position Analysis Questionnaire. Purdue Research Foundation, 1969.

11. Mitchell, J.L. and McCormick, E.J., Professional and Managerial Position Questionnaire. Purdue Research Foundation, 1980.

12. See, e.g., W.J. Phalen and R.E. Christal. Comprehensive Occupational Data Analysis Programs (CODAP). Report No. AFHRL-TR-73-5. Personnel Research Division, Lackland Air Force Base, April 1983.

* * *

FURTHER RESEARCH ON ASSESSMENT CENTERS (Symposium)

Chair: Barbara B. Penn, Port Authority of New York and New Jersey

Discussant: Terry S. McKinney, City of Phoenix

Selection of a Local City Official through an Assessment Center

Kirk O'Hara and Kevin G. Love, Central Michigan University

Assessment Centers have been adopted by many private sector industries as a primary managerial selection device, and data concerning the long-term validity of this procedure has generally been favorable (Bray, 1964; Schmitt, Noe, Meritt & Fitzgerald, 1984). Use of Assessment Centers to select local officials, however, raises unique issues for the public sector. Two such issues are:

1. gaining community support and,

2. developing and using the selection system within the budgetary range of local government.

This paper describes how these issues were addressed successfully in the application of the Assessment Center approach for the selection of Chief of Police for a small community within the State of Michigan.

Background Information

The incumbent police chief was to retire from the police department at the end of 1983. The city manager was responsible for finding a replacement for him and felt uneasy about using traditional selection procedures (e.g., resumes, interviews, etc.) to fill this important city position. He contacted the authors and conjointly they assessed the feasibility of implementing the Assessment Center procedure within the city's budgetary constraints (i.e., $3,000). It was detemined that such a procedure could be used if community volunteers could be trained as assessors and if the Assessment Center could be conducted using the facilities. It was also realized that gaining community support and input would be important to the success of the project.

Job Analysis

The development of the Assessment Center was predicted on a task-based job analysis. Although the incumbent police chief was the only individual able to provide information regarding the tasks and duties of the position, sixteen local officials and community residents were interviewed to obtain their opinions regarding the important duties and characteristics of a police chief. These interviews included the city manager, the mayor, several city commissioners, the fire chief, municipal department heads, and

the city attorney. From these interviews a list of 46 tasks was compiled. The incumbent rated each task in terms of: (1) how frequently the task is performed, (2) how difficult it is to perform, and (3) how serious the consequences are if the task is performed poorly or incorrectly. The three ratings were averaged for each task and a cutoff score established to retain only the most important job duties. Application of the cutoff yielded 34 job duties which are content analyzed into 85 knowledge, skills, and abilities (KSA's) and clustered into 11 job performance dimensions. (See Appendix A.)

Further community input was obtained by having selected community residents rate each KSA in terms of its importance to the position. A total of 23 rating forms were distributed and 18 were returned for analysis (78%). The ratings provided by community members allowed a numerical weighting (a transformation of the average KSA rating) to be applied to each of the 11 job dimensions. (See Appendix B.)

Assessment Center Exercises

The Assessment Center itself consisted of two work sample exercises and a structured panel interview. Inclusion of the panel interview allowed each of the 11 dimensions to be measured within at least two activities. (See Appendix C.) The two sample work activities were an in-basket exercise and a problem solving/presentation exercise.

The in-basket consisted of approximately 30 letters, memos, and various other paperwork likely to be encountered by a police chief, and requiring processing and disposition. The candidates were given 90 minutes to complete the exercise by writing on the memo or letter what action was to be taken. The in-basket exercise was scored independently by two trained assessors. The problem analysis/presentation exercise presented the candidate with police-oriented data that described a problem (i.e., a rise in crime rate) requiring analysis of the situation, formulation of recommendations, and a written and oral response. The oral response was presented to a simulated panel of city commissioners (i.e., assessors).

Panel Interview

The job analysis data was used to prepare a list of questions for the structured panel interview. The questions involved various aspects of police style and resolutions of hypothetical departmental problems. The panel was composed of the Chief of Police from a neighboring community and two individuals employed in the private-sector in personnel. To preserve the structured nature of the interview process, the questions were read by a fourth individual who discouraged any deviation from the predetermined format. This

allowed for a more standardized comparison between candidates. The panel members rated the candidates responses to the questions on a seven-point scale using the dimensions obtained from the job analysis.

Assessor Training

To meet the budgetary constraints of local government, in-house personnel were trained as assessors. The assessors were trained in a intensive one-day workshop, and included the assistant city manager, a city commissioner, and the local sheriff. The assessors were trained to observe and interpret effective and ineffective behaviors within each job performance dimension. Eash assessor also performed the exercises as if he/she were a candidate for the position and their performance was rated by the other assessors. In effect, the training simulated the role the assessors would play in the Assessment Center and allowed them to experience the role of the candidate as well.

Assessment Procedure

The candidates were assessed during one day using city offices and allied facilities. Seven applicants were assessed in approximately eight hours. Each candidate received a performance score on each dimension by at least two assessors. Afterwards the assessors met to discuss their ratings and to reach a consensus regarding the final Assessment Center score for the candidate on the particular dimensions they assessed. Each candidate received a score for each of the 11 dimensions. These scores were multiplied by their respective importance rating (i.e., numerical weight) and summed for each candidate to produce a final weighted Assessment Center score. Two candidates were referred to the City Manager for final selection. Reaction data from the candidates and the community indicated support and agreement with the procedure.

In summary, two of the unique issues raised in using an Assessment Center to select a city official were successfully addressed in this project. Community input and involvement was gained through the use of interviews, a mail-out survey and by training selected community residents to serve as assessors. Expenditure for the selection process was minimized by training in-house personnel as assessors and by utilizing city offices as the setting for the Assessment Center.

References

Bray, D.W. (1964). The management progress study, American Psychologist, 19, 419-429.

Schmitt, N., Noe, R.A., Meritt, R., & Fitzgerald, M.P. (1984). Validity of assessment center ratings for the prediction of performance ratings and school climate of school administrators. Journal of Applied Psychology, 69, 207-213.

Average KSA Ratings

Average Rating		
		Decision Making
4.39	1.	Knowledge of problem solving methods.
3.80	2.	Knowledge of purpose/intent of standard reports (uniform crime reports, etc.).
4.63	3.	Knowledge of current incidents/situations within the community.
4.22	4.	Skill in detecting the activities and intent of individuals.
3.75	5.	Skill in drawing inferences (conclusions) from numerical data.
4.75	6.	Skill in analyzing a situation, circumstance, or incident.
4.16	7.	Ability to comprehend and remember written materials.
4.85	8.	Ability to pursue a logical line of reasoning.
4.84	9.	Ability to reach a logical conclusion.
4.00	10.	Ability to recall details of events or reports.
		Decisiveness
4.06	11.	Skill in discriminating between messages that are significant and those that are insignificant.
4.37	12.	Skill in formulating recommendations for action.
4.33	13.	Skill in solving problems of all types.
4.85	14.	Ability to accurately assess a situation.
4.42	15.	Ability to defer a decision until required information is available.

Appendix A

Average Rating		
		Planning and Organizing
4.06	16.	Knowledge of work planning procedures.
4.12	17.	Knowledge of administrative problem solving methods.
4.01	18.	Knowledge of typical business communication procedures.
4.57	19.	Ability to organize thoughts and materials.
4.28	20.	Ability to follow established budgetary procedures.
4.50	21.	Ability to follow/develop departmental operational procedures.
4.12	22.	Ability to develop a work plan.
3.97	23.	Ability to organize departmental record keeping system.
		Political Sensitivity
4.12	24.	Knowledge of the impact of self on others.
4.12	25.	Knowledge of acceptable work standards within department.
4.37	26.	Knowledge of the roles of criminal justice system personnel (i.e. city attorney, prosecuting attorney, etc.).
4.59	27.	Skill in maintaining a professional image to the public.
4.31	28.	Ability to be objective in providing information.
4.21	29.	Ability to be objective in formulating opinions.
4.74	30.	Ability to control own emotions.
4.63	31.	Ability to exercise discretion/diplomacy in making decisions.
4.64	32.	Ability to formulate/initiate disciplinary procedures regarding departmental personnel.

Average Rating		General Police Style, Philosophy, and Knowledge
4.37	33.	Knowledge of appropriate laws and ordinances.
4.26	34.	Knowledge of common traffic problems/violations.
4.48	35.	Knowledge of standard police procedures (patrol, investigation, interrogation, etc.).
4.31	36.	Knowledge of court procedures.
3.96	37.	Knowledge of recording procedures for department (i.e. daily officer logs, report forms, etc.).
4.37	38.	Knowledge of information needs of police department (as a department and Chief of Police).
4.31	39.	Knowledge of other law enforcement agencies (local, state, federal).
4.22	40.	Knowledge of guidelines for releasing information to the public.
3.96	41.	Knowledge of typical police department structure.
4.22	42.	Knowledge of police department functions, in conjunction with other local law enforcement agency jurisdiction.
4.44	43.	Knowledge of unionized/nonunionized personnal policy guidelines.
4.22	44.	Knowledge of budgeting and fiscal record keeping procedures.
4.80	45.	Ability to provide needed service to the community.
4.65	46.	Ability to interact with all types of people.
		Oral Communication
4.44	47.	Skill in interviewing and questioning.
4.54	48.	Skill in teaching and explaining things to subordinates/others.

Average Rating		Oral Communication (continued)
4.44	49.	Skill in the oral expression of ideas.
4.60	30.	Skill in active listening to others.
4.40	31.	Ability to comprehend and express moderately complex ideas.
4.69	32.	Ability to comprehend questions.
3.99	33.	Ability to speak in public (to community groups, organizations).
4.27	34.	Ability to interact effectively in meeting situations.
		Written Communication
4.11	55.	Skill in drafting communications (letters, memos, reports, etc.).
4.22	56.	Skill in completing records accurately.
4.22	57.	Ability to communicate in writing using proper grammar, etc.
4.59	58.	Ability to communicate in writing explicit meaning and intent of communication.
		Flexibility
4.27	59.	Ability to adapt to changing situations (physically and politically).
4.55	60.	Ability to work cooperatively with other people (law enforcement personnel, city employees, etc.) under all types of circumstances.
4.69	61.	Ability to keep an open mind regarding important issues.

Average Rating		Delegating Skills
4.39	62.	Skill in evaluating the work products of subordinates (reports, handling of complaints, etc.).
4.69	63.	Ability to direct the activities of others.
3.91	64.	Ability to coordinate procedures of other departments, agencies, etc.
4.64	65.	Ability to delegate duties to subordinates.
		Interpersonal Sensitivity
4.06	66.	Knowledge of the basic foundations of human behavior (why people behave in certain ways).
3.59	67.	Knowledge of counseling techniques.
3.69	68.	Skill in counseling individuals and groups (e.g. subordinates, citizens, etc.).
4.30	69.	Skill in one-on-one interaction with others.
4.49	70.	Ability to give and receive constructive criticism.
		Leadership
3.80	71.	Knowledge of group process.
4.54	72.	Skill in group leadership (departmental leadership).
4.22	73.	Skill in conducting productive task meetings with subordinates.
4.06	74.	Ability to maintain task oriented discussion with others.
4.65	75.	Ability to motivate subordinates.

Appendix B

Dimension Ratings and Weights

Interpersonal Sensitivity: The ability to deal with a great variety of people. Listening skills, patience, sensitive to others' feelings, empathetic, and understanding of others. Tolerant of different groups and lifestyles. Ability to interact with others in a sensitive, considerate, and tactful manner.

KSA Average Rating = 4.01 Dimension Weight = 89.5%

Leadership: Provides general leadership in a variety of situations by being proactive in group situations. Commands respect of the group. Directs others to accomplishment of desired ends.

KSA Average Rating = 4.26 Dimension Weight = 95.1%

Decision Making: Uses a systematic approach to making decisions. The ability to evaluate alternative problem solutions and select an appropriate response within a reasonable amount of time.

KSA Average Rating = 4.33 Dimension Weight = 96.7%

Decisiveness: The ability to make decisions when needed. The ability and willingness to take calculated risks. The ability to make decisions under stressful conditions.

KSA Average Rating = 4.40 Dimension Weight = 98.2%

Planning and Organizing: The ability to arrange and categorize information, establish priorities, and develop a course of action designed to reach specific goals or objectives. Reviews plans of others to determine compliance to mission of the Police Department.

KSA Average Rating = 4.18 Dimension Weight = 92.9%

Political Sensitivity: Ability to identify implications of actions on others within and/or outside of the organization, and to devise positive means to gain acceptance of plans and actions.

KSA Average Rating = 4.48 Dimension Weight = 100%

Oral Communication: The ability to receive, comprehend, and disseminate information verbally. Speaks clearly, adequate vocabulary, proper grammar; clearly communicates thoughts and ideas to others.

KSA Average Rating = 4.43 Dimension Weight = 98.9%

Written Communication: The ability to convey information through written means.

KSA Average Rating = 4.27 Dimension Weight = 95.3%

Flexibility: The ability to adapt to changes in situations in order to accomplish goals/tasks.

KSA Average Rating = 4.48 Dimension Weight = 100%

Delegating Skills: Ability to distribute work to appropriate level and person for task completion, and to identify tasks that require personal responses.

KSA Average Rating = 4.40 Dimension Weight = 98.2%

General Police Style/Philosophy: The ability to view and deal with police department responsibilities in a style that emphasizes service to the citizens and community in a non-controlling mode of operation.

KSA Average Rating = 4.27 Dimension Weight = 95.3%

Appendix C

Exercises

Job Dimensions to be Assessed	In-Basket Exercise	Problem Analysis Exercise	Oral Board
1. Interpersonal Sensitivity	● ●	●	● ●
2. Leadership	●	●	
3. Decision Making	● ●	● ●	
4. Decisiveness	● ●	● ●	
5. Planning and Organizing	● ●	● ●	
6. Political Sensitivity	●	●	● ●
7. Oral Communication		● ●	● ●
8. Written Communication	● ●	● ●	
9. Flexibility	●	● ●	●
10. Delegating Skills	●	●	●
11. General Police Style/Philosophy	● ●	●	● ●

● = exercise will measure

● ● = exercise will measure exceptionally well

A Simplification of the Assessment Center Process through the Use of the Word Processor

Patrick T. Maher, Personnel and Organizations Development Consultants, Inc., La Palma, California

The assessment center process produces a comprehensive quantity of data that must be eventually placed into dimensions and scored by the raters. Research has indicated that the mere quantity of data to be considered can lead to assessor fatigue and unreliable ratings. In addition, assessor training involving the placing of behaviors is lengthy. Agencies and jurisdictions often find it difficult to adequately train the assessors due to time constraints.

This paper describes a means of eliminating or minimizing these problems by using word processing technology.

THE ASSESSMENT CENTER PROCESS

During the actual assessment center exercises, trained assessors observe and record behavior of the participants. Separate from the observation of the exercises, the assessors transcribe the recorded behavior into appropriate dimensions identified on an assessor dimension rating form.

At a time completely separate from the assessment center, the assessors meet for discussion. During assessor discussion, behavior on each participant is discussed by exercise, by dimension, and a rating is arrived at (commonly, although not exclusively, using a 5-point scale). Once all exercises have been rated across all dimensions, the assessors arrive at an overall rating for each dimension.

With some variations, assessor discussion involves the assigned assessor reading, to other assessors, his observations of behavior. The other assessors can ask questions or challenge the accuracy of the assessor's recorded behaviors. All assessors than state their score for the dimension under consideration. If there is a difference in scores among assessors, further discussion is held to determine why there is a difference. While consensus is sought, it is usually not mandated.

DEFICIENCIES

This process has a number of deficiencies. First, it is time consuming to listen to the behavior reports being read since the normal speaking rate is 120 to 180 words per minute (McCroskey, 1968). Reading summaries of recorded behaviors is considerably faster than listening to someone read the same material.

In addition, listening alone is the least effective method of obtaining information. It is difficult to understand and analyze the data, and assessors merely gain an impression and react to that impression instead of focusing on the behavior.

As time progresses, it becomes much more difficult for assessors to concentrate on the behaviors. There is a certain fatigue factor that exists merely because of the passage of time as well as the verbal environment.

Further, while the assessment center produces a considerable amount of information, the capacity of the individual's information processing system is limited and causes the assessor to employ potentially biasing heuristics to reduce the information to an amount that can be handled (Shack and Bycio, 1983).

Feedback reports prepared after the assessor discussion are made by reviewing the written observations of the assessors. On occasion, the narrative description of behavior based on assessor notes is not always consistent with the assigned ratings. These discrepancies seem to be related to the fatigue and information overload factor, which results in assessors missing important information.

In addition, many assessment centers follow a practice of having the assessors transfer their recorded observations, made during the observation of the exercises, onto a rating form listing each dimension immediately after each exercise. This practice, however, produces tremendous fatigue in the assessors as the day passes.

Further, assessors do not always place all recorded behaviors into the proper dimensions. Thus, while the behaviors are recorded in their notes, the behaviors are not always being reported and considered during assessor discussion.

COPYING ASSESSOR RATING FORMS

Initially, we remedied some of these problems by copying the assessor rating forms to which the assessors had transferred their recorded behaviors. Prior to assessor discussion, these were given to each assessor to review them and rate each dimension on a standard 1 to 5 scale. During assessor discussion, dimensions with any differences in the scores were discussed.

This accomplished several things. First, the assessors had time to review all listed behavior on all candidates prior to actually assigning any scores. Second, the assessors had some time prior to the discussion to review the material and give some preliminary assessments. Third, it avoided the extensive fatigue factor associated with lengthy assessor discussions, since actual discussion was reduced from 2 to 3 hours per candidate to less than 1 hour per candidate. Fourth, it provided a ready reference to each assessor

in trying to explain why his score was different from the others. Finally, the assessors had a better opportunity to analyze the behaviors by reading each one and giving some thought to it prior to assigning an initial score.

This process did not, however, solve the problem of making sure that all recorded behaviors were placed into proper dimensions, nor did it solve the assessor fatigue factor because the assessors were still busy transferring recorded behaviors from their notes to the rating forms.

INITIAL USE OF WORD PROCESSOR

Our next step was to take the notes of the assessors and type them into a word processor in the same order that they were recorded during the exercise (Attachment I). Thus, every recorded behavior was transferred into the word processor. Then, using the copying and delete functions, an assessor or administrator could take the recorded behaviors, copy them into each dimension, and delete the ones that did not apply to a particular dimension. This process reduced assessor workload during the assessment center by eliminating the task of immediately categorizing behavior as well as increasing the likelihood of placing behavior into appropriate dimensions.

Although this seemed to be an improvement over the oral process, the problem with this method was that while there was a listing of behaviors, there were still a few occasions in which the narrative summary of the feedback report was not entirely consistent with the score given. Further, it was possible for a recorded behavior to be eliminated as not being applicable to any dimension and therefore not considered by the assessors when, in fact, it was relevant.

CURRENT APPLICATION

The current procedure still involves the secretary typing the assessor's notes into the word processor, but now the entire list is printed. The administrator then reviews all of the behaviors and dictates a summary in each relevant dimension. Assessors are then given a copy of the original list of recorded observations of behavior and the narrative summary (Attachment II).

Those assigned as the primary assessor in a given exercise have primary responsibility for reviewing the summary for accuracy and consistency against the recorded behaviors. All assessors are also charged with a secondary responsibility of reviewing all behaviors of candidates not assigned to them with the same intent. This serves as a check and balance by having others verify the accuracy of the narrative summary. Discrepancies are noted during assessor discussion and corrected before a final score is assigned by assessors.

Further, when one person summarizes all candidates, he can note overlapping factors or differences in performance that could be added to the summary. As before, the assessors rate each dimension prior to the assessor discussion, and if there is any difference in scores, assessors attempt to resolve them.

This process seemed to solve most problems outlined above, with one exception. We followed a process wherein each dimension was discussed for each exercise and rated. At the end, we reviewed ratings by exercise and arrived at an overall rating for the dimension. If the ratings were different for different exercises (e.g., oral communication having a 2, 3, and 4 for the 3 different exercises) then assessors had to arrive at an overall rating. While told not to average scores, but rather to consider several factors including the extent to which an exercise actually required a dimension to be shown, some averaging seemed to be taking place. In following the process outlined above, it was difficult to review all behaviors across the exercises because the data was spread over several pages. It seemed that the inconsistencies between feedback reports and some scores could be traced to this problem. That is, individually in an exercise, a behavior ("strong" or "weak") did not significantly impact the individual rating within an exercise. However, because the feedback reports placed all behaviors (or their summaries) into one continuum, then the several behaviors combined would impact the overall rating.

To solve this problem, we changed the report format. While we still discuss the dimensions by exercise and rate each dimension separately by exercise before assigning an overall score, we list the summary of each exercise within the dimension. Thus, when arriving at an overall score for a dimension, the assessors can readily read through the entire summary and base their overall score on behaviors rather than mere numbers.

In the initial assessment centers, we held assessor discussion fairly soon after the assessment center. In most cases it was only a few days, and never more than a week. Assessors had the ability to recall individuals fairly easily.

Now, we allow more lead time to prepare for assessor discussion. A secretary must first transcribe the notes. Then an assessor must review the behaviors and dictate a summary of them. Next, all material and the original notes must be returned to the assessors, who review them for accuracy and rate each candidate. We are faced with up to two weeks of lead time (although this can be reduced). Assessors are now finding it extremely difficult to recall individuals, especially where we have more than six candidates spread over 2 or more days. There has been some discussion about providing pictures so that assessors can recall individuals, but this has not been implemented. While there have been occasions where recalling the individual has provided additional information about the individual, this was more prevalent in the earlier process, where the documentation

was not as complete. Our feeling is that if the assessor does not recall the individual, he is basing a rating of performance on the available behavior and not on other subjective factors.

This, unfortunately, also causes another problem in that assessors do not gain the experience or training necessary to properly place behaviors into dimensions. Thus, only a few individuals gain this ability. Where assessors are used on a reciprocal basis and often for only one time by a given agency, this is not a problem. However, where assessors are used repeatedly, they do not develop this skill unless they are assigned to dictate the narrative summary for a certain number of participants in each assessment center, which also reduces the burden on the administrator.

SUMMARY

Our experience has shown that the use of the word processor reduces assessor fatigue, and is more economical and superior to the more common method of reading behavior.

ATTACHMENT I

Following is a partial printout based on an assessor's notes from a non-assigned role group discussion. The complete copy is used to dictate the summary report and is provided to all assessors so that they can review all recorded behavior.

I think we should...
Want to use a stop watch?
Excuse me (takes off coat)
Should use board...
Keep notes
Your question Scott
Not to interrupt but we determined physical fitness test of strength
Need to prioritize
Pursue 2 alternatives--Heart is one
Back and muscles are another problem
I think so, need to do so for legal reasons
Eye contact with others as they/he speak
(summarizes quest from Jim)
We're getting into 1/2 hour now

The following is a partial printout of the participant's written responses to in-basket items. The item number is indicated and the verbatim wording of the participant is indicated for each item. Assessor comments are enclosed in parentheses. It too is used to dictate the narrative report and is also provided to all assessors so that they can see the entire in-basket activity of the participant. Assessors have a copy of the in-basket so that they can review the item to determine the context of action taken on individual items.

1 Bruington: Capt. Bruington, please attend the staff meeting this Tuesday. I will be out of town and unable to attend. Thank you. I. M. Brite

2 Bruington: Capt. Bruington please handle this matter. I suggest that we peruse this through legal channels. If Sanders is guilty then the Department will take necessary disciplinary action. Thank you. I. M. Brite PS. was this on duty. Reply on 6-16. Thanks.

3 Sharon: Thank you for your messages. Please inform Mr. Robingson I will call and make an appointment upon my return. Also, can you provide me with information about Mr. Robinson. Thanks. I. M. Brite

4 Information only. (Noted that Papik was in charge of recruit training and that Bruington was in charge of in-service training and related matters).

5 Bruington: Capt. please reply to DC Harman in regards to his attached request. Have a copy of your reply for my information on this matter. I would appreciate this being handled by 6-20. Thank you. I. M. Brite

ATTACHMENT II

Following is a partial narrative report prepared from behaviors printed from the report dictated from assessor notes or the written product of the participant. A summary of each exercise is provided, followed by each dimension which contains a separate paragraph on each exercise being rated. This report is provided to all assessors and is used by them to make initial ratings prior to assesor discussion. After assesor discussion, the report is printed in final format, with any changes made during assessor discussion, and serves as the feedback report for the participant.

EXERCISE SUMMARY

In the written problem, Pat wrote a nine-page report. Pat did not feel that it was a difficult problem, and felt that the easiest thing for him was to obtain high points for discussion since this could easily be accomplished by the use of a highliter. His overall strategy was to balance his time between the divisions of the report, and he saw his major priority as the recommendation of a policy.

ORAL COMMUNICATION SKILL

In the group discussion, Pat was an effective communicator, expressing his ideas clearly and loudly to the group, although he had a tendency to be somewhat stiff and he lacked animation. He was very active verbally. He was cut off on a few occasions, possibly due in part to his failure to use communication techniques (e.g. standing, gesturing, etc.), that may have enabled him to gain or keep the floor. Overall, however, Pat was very effective in getting his ideas or thoughts out to the others.

WRITTEN COMMUNICATION

In the in-basket, Pat had two spelling errors, six punctuation errors, one sentence fragment, one run-on sentence, and one sentence that was awkward in its construction. Generally, Pat's memos were clear and concise, and there was little or no problem in understanding what he was trying to say.

In the written problem, Pat had four spelling errors, ten errors in punctuation, two errors in the use of the plural, one sentence fragment, three awkward sentences, and he omitted a word once. Pat wrote in a straight narrative style with no headings, but his report was logical and easy to follow. Headings, especially given the length of his report, would have aided the reader in following the report and in referring to other portions for clarification and review.

DEVELOPMENT OF SUBORDINATES

In the group discussion, Pat identified the need to pursue two alternatives in the physical fitness program, the heart or cardio-vascular program being one, and back and muscles being the other major problem. Pat also indicated that the priority should be based upon identifying individuals who are not physically fit before they were lost due to injuries. Pat also noted that if the department could not sell the personnel on the importance of maintaining good health, then the program was going to become punitive in nature. He also wanted the group to decide how they were going the deal with those individuals who failed the program and indicated that at some point negative discipline was going to be necessary.

In the in-basket, Pat made full use of his staff in handling the items. Pat started off by asking that his subordinates provide him feedback, but towards the end he omitted this, apparently due to the fact that he was running short of time, and some items were being handled in a little more superficial manner because of the lack of time remaining.

REFERENCES

Miller, Wilma H., Untitled. New York, N.Y.: Center for Applied Research in Education, 1973.

McCroskey, James C. An Introduction to Rhetorical Communication. Englewood Cliffs, NJ: Prentice-Hall, Inc, 1968.

Shaw, Mitchell S. & Byrd, Peter. The Assessment Center as Psychological Process: An Analysis and Recommendations. Paper presented at the annual meeting of the International Personnel Management Association Assessment Council, Washington, D. C., May, 1983.

Standards and Ethical Considerations for Assessment Center Operations, December, 1978, Task force on Assessment Center Standards.

* * *

Replicating Research on Police Promotional Assessment Centers

Dennis Joiner, Dennis A. Joiner and Associates, Sacramento, California; and Phil A. Carlin, City of Tucson, Arizona

Research is presented on the issue of integrated versus unintegrated data (consensus judgement versus mathematical combination of scores) to determine the final results in a police promotional assessment center. A 1982 study produced correlations ranging from .89 to .99 depending on the specific comparisons computed. This paper presents the results of two different replications of the study. In each replication different factors were varied to isolate the cause of the high correlations obtained in the prior study.

Working from statistically derived job analysis results, supplemented by situational data and work samples, four exercises were developed which would simulate the most essential task areas in the classification of Police Lieutenant. This would allow assessors to observe, record, classify and evaluate job relevant behavior in job relevant situations. Using job simulation exercises tailored specifically to the classification of Police Lieutenant as used in Tucson would not only increase candidate acceptance and compliance with legal requirements for content validity, but also allow candidates to "get into" the simulations "as if" they were real life.

The exercises developed include:

- Leaderless Group Discussion: In this exercise, candidates in groups of six or seven were given a number of current issues and problems confronting the department and were instructed to formulate specific recommendations or decisions for dealing with each. The group interaction was observed by the assessor team, each assessor paying particular attention to two or three assigned candidates.

- Oral Presentation Exercises: This exercise took the form of a management meeting. Candidates were allowed a brief time to plan, organize and prepare a presentation on an assigned topic to their supervisors. They would then present their ideas and respond to questions. The assessors played the role of the Captain and the Deputy Chief of Police and asked the candidate to respond to a series of pre-determined (standardized) questions.

- In-Basket Exercise: This exercise consisted of a variety of materials of varying importance and priority which would typically be handled by an incumbent of the classification. Candidates were given a limited amount of time to deal with these materials. They were later interviewed by assessors who reviewed with the candidates how they handled the material and their reasoning in doing so.

- Written Report Exercise: Candidates, in this exercise, were given a job relevant topic pertinent to the position and were instructed to provide a written report. The written document was later received and rated independently by two assessors.

Once the exercises had been developed, the consultant again visited Tucson. During this visit, all the developed exercise materials were discussed with the agency selection specialist and top Police Department management in relation to the supporting documentation from the job analysis.

Administration

Candidate orientation is a very important part of any assessment center examination process. In fact, in the authors' experience, the few protests that are filed on this type of examination are based on a lack of knowledge of the process and suspicious which result from lack of knowledge. This approach to candidate orientation was to send general information to candidates in written form. In addition to the written material, all candidates reported for a two hour orientation session prior to their participation in the first examination test instrument or exercise.

Even though all individuals selected to serve as assessors had prior training and experience in evaluating candidates at this level, additional training was necessary due primarily to the custom nature of the examination and the need to standardize scoring tendencies within the specific group which was assembled. Prior to the on-site assessor training, each assessor received a comprehensive package of pre-reading materials related to the specific assessment job at hand.

During the process, each candidate was independently observed and evaluated by two different assessors in each exercise. The process was scheduled such that upon completion of the exercises, each candidate had been evaluated once by each assessor. It was necessary to obtain eight assessors for the four exercises. Because the report writing exercise involved making recommendations to improve the community relations and crime prevention programs of the Tucson Police Department, two of the eight assessors were recruited from other law enforcement agencies in the Tucson area. These two assessors, who knew the local environment and existing programs of the Department, evaluated the written reports submitted by all 52 candidates. (These reports were blind rated to ensure that no halo effect would occur through name recognition.) The other six assessors were selected from agencies outside of the state of Arizona. These six assessors, who had no prior knowledge of the candidates or access to background information regarding the group, observed, recorded and evaluated the effectiveness of candidates in the remaining three exercises.

The procedure for establishing candidate ranking was a combination of cumulative scores on each performance dimension with the addition of an overall evaluation score which the assessors established by consensus. The nine performance dimensions were individually weighted on the extent to which they differentiated between levels of effective performance on the job. The overall consensus score was allocated the weight of an additional 10 percent or in effect a true 9.09 percent of each participant's final score. This overall evaluation score (often referred to as the Overall Rating or OAR) allowed assessors to assign a score based on a consideration of the total picture of the candidate which was generated by discussing the candidate's performance in all four exercises.

Administration Time Requirements

The total candidate group of 52 candidates required one full day of assessor training, four full days of assessment (13 candidates per day) and two days of post-assessment evaluation of candidate performance by the assessors to develop the final ranked list.

Results

The consensus of the assessment team was that the top 25 of the 52 participating candidates on the score ordered list had demonstrated sufficient skills to be considered job ready and placed on the eligible list. This recommendation was accepted by the Civil Service Commission and the list was adopted during the week following the examination. No protests or appeals were voiced by the candidates regarding the content, methods or procedures used in the examination process. The first eight appointments from the list were made immediately following its adoption.

Research Conducted

To investigate the impact of the final evaluation sessions on the final results of this examination, the authors were committed to seeing that all rating forms completed independently by the assessors immediately after the exercises were photocopied prior to the integration session. These rating forms which included each assessor's initial-tentative scores were stored for later research.

Weeks after the assessors had met over a two-day period to discuss and finalize all scores for all dimensions and assigned overall consensus scores for all candidates, the initial ratings were compiled without the overall consensus scores. Three different correlations were then computed. First, a correlation coefficient was generated to assess the impact of the overall consensus score which had the weight of an additional 10 percent on the integrated dimension scores: correlation coefficient = .908174 when correlated with final results, .890058 when correlated with integrated behavior dimension scores minus the overall score. These correlations indicate that there might have been a difference in the rank order list if the overall consensus score had been the only score used

for rank candidates. In fact, inspection of the actual rank order list produced by the examination and a simulated list produced solely on the basis of the overall consensus scores shows some movement of candidates. The biggest impact, however, of ranking by overall scores is a "grouping" of candidates into a series of ties (52 candidates into 12 groups).

But, the overall consensus score is rarely used in the public sector as the sole determinate of rank on an eligible list and in this exam had an actual weight of 9.0909 percent. The overall consensus score was intended only to round out the rough edges produced by the mechanical combination of the dimension scores.

A second comparison was made between the integrated dimension scores (without the overall rating) and the final scores (with the overall rating): correlation coefficient= .998851. Looking at the rank order list of candidates produced with and without the overall score produces no change in the order list of candidates.

Finally, a comparison of the pre-integration dimension scores and the post-integration scores (including the overall score) produced a correlation coefficient of .983771. The rank order lists produced by these data are also identical suggesting that at least in this case, the integration session made no difference in the final results of the examination.

The Need for Replication

Two major hypotheses were developed to explain the high correlations obtained when comparing pre-integration and post-integration session scores. The first was that the prior training and experience of the assessors caused the consistency of the pre and post data; i.e., the data (scores) were accurate to begin with so no or little change was necessary in the integration session. The second had to do with the administration model. More specifically, since all 52 candidates were assessed prior to the integration sessions (13 candidates per day for 4 days straight), it was proposed that even with the polaroid photos of the candidates which were used in the integration session the assessors may not have remembered the candidates' performance well enough to be comfortable moving scores very far from the tentative scores assigned while the information was fresh.

PART TWO: REPLICATION WITH THE 1984 POLICE LIEUTENANT ASSESSMENT CENTER

The Replication Process

By the summer of 1984 fifteen Police Lieutenants had been appointed from the eligible list produced in 1982 when the list expired. The same consultant was hired to conduct an update of the job analysis,

prepare and administer another assessment center to establish a new rank-ordered eligible list.

The assessors selected for this assessment center were again all from outside the Department but, unlike the 1982 Lieutenant exam, the assessor team consisted of half individuals with prior training and experience with the assessment center process and half with no prior experience. This would allow for comparison of pre-integration and post-integration scores of experienced and novice assessors.

Research Results

The correlation coefficient produced for the total population (N=39) when correlating the overall consensus score with the final score is .960053 (1982 Lt. = .908174).

Comparing the integrated dimension scores (without the overall rating) and the final scores (with the overall rating): correlation coefficient = .998806 (1982 Lt. = .997751) which indicates that the overall rating had very little impact on the final scores.

Comparing pre-integration dimension scores and the post-integration dimension scores produces a correlation coefficient of .988754. This correlation indicated that, as in 1982, the integration process had very little impact on the final scores.

Comparing pre-integration to post-integration dimension scores for the first two days of the 1984 Lieutenant examination (N=20): correlation coefficient = .971169. Making the same comparison for the second cycle of the 1984 exam (N=19) produces a correlation coefficient of .996504. A slight improvement in consistency. Due to the high correlations obtained and slight differences between the first and second cycles no further research was conducted comparing the novice to more experienced assessors.

PART THREE: REPLICATION WITH THE 1984 POLICE SERGEANT ASSESSMENT CENTER

Research Results

Comparing overall consensus score with final score: correlation coefficients = .922 (82 Lt. = .908; 84 Lt. = .960)

Comparing integrated dimension scores with the post-integration dimension scores: correlation coefficients = .997 (82 Lt. = .998; 84 Lt. = .999).

Comparing pre-integration dimension scores with the post-integration dimension scores: correlation coefficients = .977 (84 Lt. = .989).

All correlations are quite consistent with the high correlations obtained in the two Police Lieutenant assessment centers and would lead one to question the value added by the integration session.

Conclusions

Prior to drawing any final conclusions the authors looked at the rank order lists produced for the 1984 Lieutenant and Sergeant examinations and compared them to the simulated lists which would have been produced by the unintegrated scores. The chart below shows the number of promotional candidates who changed rank order positions when moving from the unintegrated to integrated data produced lists and illustrates the number of rank-order positions of change.

1984 Lieutenant		1984 Sergeant	
Number of Candidates	Movement in Positions On List	Number of Candidates	Movement in Positions On List
16	0	10	0
15	1	10	1
5	2	7	2
2	3	2	3
1	4	8	4
		1	5
		2	6
N=39		2	7
		1	8
		N=43	

The impact of the integration session on individual candidates can be great even when the correlation coefficients between pre and post integration are quite high. These rank order differences are even more significant when you consider the number of agencies which operate from a tradition of promoting from number one down on the rank-order list. This practice is particularly prevalent for public safety classifications which are the classifications for which assessment centers are most often used in the public sector.

The greatest limitations of these data are the small sample sizes used. The authors are quite interested in furthering this research. If you are interested in replicating this research please contact either author. We will compute the correlations if provided with the raw data.

EXERCISE CORRELATIONS

1982 Police Lieutenant

Unintegrated

	L.G.	I.B.	O.P.	R.W.
L.G.	-	.47	.31	.45
I.B.		-	.18	.14
O.P.			-	.24
R.W.				-

Integrated

	L.G.	I.B.	O.P.	R.W.
L.G.	-	.51	.33	.43
I.B.		-	.18	.14
O.P.			-	.24
R.W.				-

1984 Police Lieutenant

Unintegrated

	L.G.	I.B.	O.P.	R.W.
L.G.	-	.06	.31	.11
I.B.		-	.36	.01
O.P.			-	.13
R.W.				-

Integrated

	L.G.	I.B.	O.P.	R.W.
L.G.	-	.06	.32	.02
I.B.		-	.38	.01
O.P.			-	.09
R.W.				-

1984 Police Sergeant

Unintegrated

	L.G.	I.B.	R.W.	P.O.
L.G.	-	.30	.01	.06
R.P.		-	.11	.18
R.W.			-	.07
P.O.				-

Integrated

	L.G.	I.B.	R.W.	P.O.
L.G.	-	.28	.04	.11
R.P.		-	.14	.21
R.W.			-	.05
P.O.				-

L.G. = Leaderless Group
I.B. = Inbasket Exercise
O.P. = Oral Presentation
R.W. = Report Writing
R.P. = Role Play
P.O. = Patrol Operations Problem

* * *

VALIDATION, IMPLEMENTATION, TRANSPORTABILITY AND UTILITY OF A SELECTION PROCEDURE FOR PROFESSIONAL CLASSES IN A STATE MERIT SYSTEM (Symposium)

Chair: William Rowe, State of Louisiana

Validation of the Professional Entrance Test (PET) for the State of Louisiana

David A. Dye, Psychological Services, Inc., Washington, D.C.

PSI was under contract to provide four deliverables to the State: a job analysis report, a survey of options report, a validation report, and a testing manual.

A comprehensive job analysis was conducted on 19 professional occupations, encompassing 46 classifications. The purposes of the job analysis were to: identify important work behaviors and employee competencies; group occupations into job families; and, serve as a basis for criterion development. With this in mind, the key toward its development was to come up with statements that would apply across the broad range of occupations and still adequately cover the critical and important aspects of their classifications. PSI, in working with State personnel staff, generated lists of work behaviors and competencies, reviewed them, and put them into a final questionnaire.

THE JOB ANALYSIS QUESTIONNAIRE

Administration

The questionnaire was then administered by trained State personnel to incumbents between January and March, 1983. A total of 2,835 questionnaires were completed and forwarded to PSI for analysis.

Analysis

In addition to identifying critical and important work behaviors and competencies, I mentioned that the job analysis served another purpose—grouping of occupations into job families. Since it would have been too costly and undesirable to conduct a separate validation study for each of the 46 classifications, we decided to use the job analysis ratings to group the classes into job families based on the importance of their work behaviors. This was performed statistically by means of a cluster analysis. Our intent was to come up with a set of families that would satisfy the desirable properties of the cluster analysis, but would also meet sample size requirements for conducting criterion-related studies. The clustering scheme which seemed to make the most sense resulted in identifying four job families: facilitative,

research and investigative, technical and administrative, and determinations (disability examiner and program analyst). The number of classifications in the job families ranges from four to seventeen. The total number of incumbents ranges from 48 to 2,541.

Summary of Options Survey

As presented at the 1983 IPMAAC conference, PSI reported on its efforts to locate an existing cognitive abilities test that would be suitable for entry-level selection into the 19 professional occupations. If the State could use an existing test, certainly this would forego the expenses associated with test development and thereby speed up the validation study. Well, from an extensive search that we conducted involving questionnaires sent to other state governments and IPMA test users as well as a search of the literature and based on conversations with other major test publishers, PSI concluded that no single, fully appropriate test battery could be found. State personnel agreed. Thus, PSI received the go-ahead from the State to develop a test suitable for selection of personnel into highly cognitive, complex jobs.

Summary of Test Development

Much of our thinking into the development of actual test items was based on what has worked well in the past. Previous research had shown a variety of item types that would be appropriate for our purposes. We chose four item types--tabular completion, inference, reading comprehension, and quantitative reasoning.

Using the results of an item analysis and an item fairness analysis which detects items biased toward any one group, we then selected the best mix of questions with high item-total correlations and appropriate difficulty levels. The final test consisted of 100 questions: 25 of each item type.

VALIDATION

Our next step was to validate the test. You will recall that we performed a cluster analysis and that we had intentions of linking the testing procedures into the cluster analysis for the purpose of selecting occupations on which to do criterion-related studies. There were additional considerations in conducting criterion-related studies--technical and administrative feasibility.

Our first consideration in identifying occupations was sample size. We figured we needed a sample size of at least 200 for adequate statistical power, and should deviate from this only if necessary. It turns out that of the 19 occupations, only 4 met the sample size requirement. They were Eligibility Worker, Probation and Parole Specialist, Human Services Worker, and Employment Security Interviewer. We also decided to include

Computer Programmer/Analyst because previous research (Schmidt et al., 1980) had shown that we might have adequate power with a smaller sample.

Now that we have satisfied our technical requirements, it was time to face the real world challenge of "can we really administer it to these occupations?" Specifically, would enough research participants be available for testing and was it possible to develop criterion measures other than supervisory ratings? In meeting with State SMEs, it was determined that two of the occupations, Employment Security Interviewer and Human Services Worker did not lend themselves to criterion development.

So, we were left with three occupations. Eligibility Worker and Probation and Parole Specialists were the most populous occupations in their respective job families. These job families accounted for 13 of the 19 occupations and about 86% of all incumbents. By including Computer Programmer/Analyst we brought the totals to 17 of the 19 occupations and 99% of the incumbents. Most importantly if validation studies were conducted on these three occupations, we would have essentially "covered" all but two occupations, those being in the Determinations job family. Dick will have more to say on what I mean by "covered" when he speaks on the method he devised for transporting validity to the other occupations. A key factor in the success of this project, and very likely, one of main reasons for achieving the results that we did was the special care taken in constructing the criterion instruments. Certainly, validity generalization research has shown the impact that poorly constructed, unreliable measures of job performance can have on the validities we obtain. Through a series of introductory meetings, pretesting, and review sessions, a combination of special ratings and work samples tailored to each occupation were developed. Additionally, a job knowledge test was devised for the Eligibility Worker occupation.

To take an example of one of the work samples for Eligibility Worker, one exercise requires a set of supporting documents, and written narratives to three specific questions in order to determine a person's eligibility to receive food stamps.

With the final criterion measures in hand, it was time to administer them and the PET to incumbents in the 3 occupations. Between May and June, 1983, data were collected on 589 incumbents from 11 locations throughout the State. Employees spent the better part of a day answering test questions and performing the work samples. All of the data was forwarded to PSI, scored, and keypunched or optically scanned. As a 100-item test, it is not a particularly difficult test, but with a standard deviation of about 16, it still provided a sufficient range of scores to enable accurate measurement of persons with high levels of cognitive ability. The subtest reliabilities were calculated by means of the KR-20 formula; the total test a linear composite of the subtest reliabilities. All reliabilities are moderate to high.

In addition to the criterion scores collected, composite criterion scores were computed by first standardizing the individual measures and then adding them together. This was done to give equal weight to each of the components.

For all occupations, reliability of the performance ratings was estimated by correlating the average of the ratings with a separate rating of overall performance. However, the most appropriate method for estimating reliability of supervisory performance ratings is the correlation between ratings separated by a time interval of several months and made by separate raters. Reliability calculated in this way is typically about .60. The reliabilities we obtained ranged from the high .70s to the high .80s. The effect, then, of our method is to overestimate reliability estimation due to the open-ended questions; and therefore, reliability was not computed for Eligibility Worker and Probation/Parole Specialist. It was computed for Computer Programmer/Analyst by correlating performance on odd- and even-numbered questions, and correcting it by Spearman-Brown. Its estimate was high at .91. Likewise, the only composite for which reliability could be calculated was Computer Programmer. It, of course, turned out to be very high at .95. Nevertheless, the criterion measures were judged to be adequate for continuing with the validity analysis. While some of the criterion variances were low and some of the reliability estimates were lacking, it was recognized that the overall effect would be to underestimate the true validity of the PET.

For the validity analysis, Pearson product-moment correlations were computed for each of the three occupations between the PET and the applicable criterion measures. The resulting validity coefficients provide strong support for the validity of the PET. Furthermore, the coefficients obtained in this study, if fully corrected for statistical artifacts, would be consistent with the findings of Hunter's research on the General Aptitude Test Battery; specifically that the true validity of cognitive ability test for jobs of high complexity is about .56.

In addition to the validity study, and in conforming with the requirements of the Uniform Guidelines on Employee Selection Procedures a fairness study of the PET was performed for the occupation of Eligibility Worker on the basis of race. Eligibility Worker represented the only occupation in which there was a sufficient number of subgroups to perform such an analysis, 116 whites and 99 blacks. In no occupation was there an adequate sized sample for looking at sex differences.

Analyses performed with the Wilks-Gulliksen procedure (a three-step, sequential comparison of the black-white regression line differences, in terms of standard errors, slopes, and intercepts) suggest no difference between the black-white regression lines. If the common regression line (based on both groups) were to be

used, it would result in overprediction of blacks' job knowledge test performance. The PET, in this instance, favors blacks and is somewhat unfair to whites, despite the fact that whites scored higher on the test. This difference, though, would appear to have little effect. The conclusion of the fairness analysis is thus consistent with the literature results; it is not unfair to blacks. The overall findings then are that the PET is valid and fair to all groups.

PSI wanted to investigate the possibility of shortening the test, making it more practicable, but still maintain its high degree of reliability and validity. At the same time, wouldn't it be nice to come up with parallel forms : Form A and B? The plan was to develop two tests of 40 items, each containing 10 items from each subtest.

First, let's look at reliability. Reliability for the original version calculated by the KR-20 formula gave an estimate of .93. By replacing the PET with a shortened version that is 4/10 in length, the Spearman-Brown formula estimates a decrease in reliability to .84. Certainly, still respectable. On the validity side, the original version shows a validity of .42. This figure represents the average validity in predicting the composite criterion of all three occupations, not corrected for any statistical artifacts. Using the formula to predict the change in validity by decreasing the number of items from 100 to 40, the expected validity is .40. Not much change. So, with the chance of being able to develop and market two tests from an existing test with little loss in reliability and virtually no loss in validity, and providing users with a test that takes 60% less administration time, the shortened version showed great promise.

Taking the 100 existing items of the original version, the shortened version was constructed by choosing moderately difficult items with high reliability; that is, those with high positive item-total correlations. Also, items with inefficient distracters and those suspected of being biased against blacks were not used. Consideration was also given to balancing item content and the distribution of keyed alternatives. The parallel forms were constructed then to be matched with respect to difficulty and discriminability.

As a check on the accuracy of the formulas, the short form statistics were then calculated by rescoring the tests with the new items and correlating performance on the test with the composite criteria.

* * *

USE OF VIDEO IN ASSESSMENT (Video Presentation)

Chair: Nancy Whitlock, National Passenger Railroad Corporation, Washington, D.C.

Orientation to Assessment Centers - A Video Approach

Dennis A. Joiner, Dennis A. Joiner and Associates, Sacramento, California

Candidate orientation is a critical requirement for the success of any assessment center examination process. It was not long ago that it was important to provide an orientation to potential assessment center candidates because no one in the candidate pool had heard of the process. In recent years it has become more important to provide a thorough orientation because some of the candidates have heard of and participated in assessment centers as either a candidate or assessor in your jurisdiction or a neighboring organization. It then becomes important to equalize the knowledge of the assessment center process within the candidate group. This will reduce any possible unfair advantage or perceptions that someone had an unfair advantage due to prior exposure or different levels of familiarity with the technology.

The primary purpose of candidate orientation, however, remains the same: to reduce artificial test stress (test anxiety) in order to allow candidates the opportunity to demonstrate a truer level of their job related skills, thus improving the predictive validity or accuracy of the examination.

There are a number of different methods for providing candidates information in advance of their participation in an assessment center. These can be grouped into three general modes: 1) written, 2) oral or classroom, and 3) video. These methods are not mutually exclusive and utilization of all three is highly desirable where logistically feasible.

Written information can include all of the essential details such as: the purpose of the program, how the information will be used, a general description of the process, opportunity for performance feedback and how it will be provided, when and where to report, etc. Other more general information such as articles describing applicants of the assessment center process can also be provided to participants well in advance of their participation. The advantage of written material is that all potential participants, regardless of the size of the candidate group, receive the same information. The disadvantage is the lack of opportunity to ask questions and actually see what an assessment center process is like.

Oral and classroom orientation can take many forms, ranging from a brief question and answer session at the examination site to participation in selected parallel forms of the exercises to be used or a simulation of the entire assessment center process or a mock assessment center.

Mock assessment centers are usually only practical with promotional groups. The brief overview and question and answer session on the morning of the exam is much more common when candidates are coming in from outside of the jurisdiction.

Use of videotapes for assessment center orientation are becoming increasingly popular for a number of reasons including: 1) A videotape provides a standardized approach to orientation which can be viewed by candidates in different locations at the same or at different times; 2) Videotaped examples of exercises allow individuals to actually see what assessment center exercises are like, and 3) Videotaped orientation material can also be used for orienting user department managers and novice assessors during their training.

The 1985 videotape "Assessment Centers: What Are They?" was shown for those in attendance. This 50 minute videotape written and produced by Dennis and Sherry Joiner allows the viewer to follow a group of six candidates through an orientation session and four common job simulation exercises (a leaderless group discussion, an inbasket exercise with an interview, an oral presentation and two examples of a role play or subordinate counseling exercise). The various segments are linked together with commentary explaining the rationale behind each type of exercise and why assessment centers are such valuable tools for making selection, promotion and career development decisions.

* * *

IPMAAC PROFESSIONAL AFFAIRS COMMITTEE FORUM

Professional Ethics: Requirements, Issues and Practicalities

Chair: Jurutha D. Brown, City of Los Angeles

Participants: Marilyn K. Quaintance, Morris and McDaniel, Inc.; William B. Owen, U.S. Department of State; and, Jennifer French, San Bernardino County, California

Discussant: Glenn McClung, Denver Career Service Authority, Denver, Colorado

The 1984-85 Professional Affairs Committee sponsored a forum on "Professional Ethics: Requirements, Issues, and Practicalities". Marilyn Quaintance, Chair of the Committee, provided an overview of ethical requirements for public personnel assessment professionals.

Bill Owen summarized ethical requirements contained in the newly published Standards for Educational and Psychological Testing and Jennifer French described ethical dilemmas confronted by public assessment professionals in a municipal setting. Glenn McClung served as discussant.

Dr. Quaintance began by defining "ethics" as "the study of standards of conduct" and moral judgement and "the code of morals of a particular profession". "Moral" was defined as "capable of making the distinction between right and wrong in conduct or behavior", with "morals" being a set of principles, standards or habits with respect to right or wrong conduct. Finally, "ethical" was defined as conformity with an elaborated ideal code of moral principles, sometimes specifically with the code of a particular profession". Dr. Quaintance suggested that when a set of principles or a set of ethics is adopted that the profession view these standards as an ideal toward which to strive.

The six stages of development of moral judgement, identified by Kohlberg, were presented. The first stage, Punishment Orientation, which involves obeying rules to avoid punishment, was contrasted with the final stage, Ethical Principle Orientation, in which actions are guided by self-chosen ethical principles.

Ethics was presented as a two step process - an educational process (i.e., helping others to understand right from wrong and to reason about moral judgement) and an enforcement process. The first process of defining and understanding standards was the major focus of this presentation. Dr. Quaintance reviewed the documents containing ethical provisions that applied to the profession of the public personnel assessment. These included the IPMA Code of Ethics, the IPMAAC Code of Professional Principles, the Standards for Educational and Psychological Testing, The Division 14 Principles for the Validation and Use of Personnel Selection Procedures, the Uniform Guidelines on Employee Selection Procedures, the Standards for Providers of Psychological Services, the Speciality Guidelines for the Delivery of Services: Industrial/Organizational Psychologists, and the Ethical Principles of Psychologists.

Dr. Quaintance concluded by emphasizing that there is a continuing need for the IPMAAC membership to communicate these ethical provisions to new members entering the profession and to emphasize the educational process through the publication of articles designed to aid others in understanding and complying with ethical requirements.

Mr. Owen started his presentation by emphasizing that "professional is ethical". If you are being professional, you are being ethical, and if you are not being ethical, you are not being professional. He indicated that the Standards for Educational and Psychological Testing, that provide some guidance as to ethical behavior, took years to develop and probably will take more years to interpret.

Mr. Owen provided an outline of the Standards. The Standards contain four sections: 1. Technical standards for test construction and evaluation; 2. Professional standards for test use; 3. Standards for particular applicants; and 4. Standards for administrative procedures. The first section on technical standards deals with validity, reliability, test development, scaling, norming, meta analysis, etc. The professional standards for test-use cover applicants of testing including clinical, educational, counseling, employment, licensing, certification, etc. The particular applicants cover linguistic minorities, the handicapped etc. The final section on administrative procedures has guidance on test administration, scoring, reporting and test-taking rights.

Mr. Owen stated that it is clearly unethical if we use a test when we know that the test is not appropriate for a given purpose. Further, Mr. Owen indicated that control of the test administration situation so that it is standardized and so that the test results are meaningful is an ethical obligation. Finally, Mr. Owen summarized the ethical responsibilities of test users to test takers. These three areas - the use of tests, the administration of tests and the rights of the test takers were felt to be areas in which the Standards provide guidance to professional practitioners.

Ms. French's presentation focused on the practicalities of implementing ethical requirements in San Bernardino County, California. She stated that eight professional test development professionals had responsibility for 300 recruitments per year. This is approximately 55½ hours on each test development effort. Ms. French suggested that, when confronted with this workload, it was virtually impossible to meet all standards and guidelines. In contending with that reality, she emphasized the importance of allocating resources in a manner appropriate to the criticality or potential impact of each assignment. Ms. French stated that, nevertheless, the existence of the technical and ethical standards made the test development staff do a much more professional job. She said, "They have made us all much more aware of the very most important aspect of the work we do. They have made us better practitioners." Ms. French stated that the standards have stimulated research and made for interesting discussions, debates and arguments. They have brought us together in our professional organizations, strengthening those organizations and the profession.

While the ethical standards seem to suggest situations that are "black" or "white", in actuality they are "gray" calling for professional judgement. Ms. French gave specific examples of the ethical situations confronting her in her role with municipal government.

Mr. McClung focused on defining the interrelationships of public assessment professionals - with each other, with management and with politicians and the conflicting demands of the many diverse groups on us. He stressed that standards and guidelines have helped to pull our profession together. In particular, we have been united by our efforts to make those standards more reflective of the real world.

* * *

USE OF RATINGS AND SELF ASSESSMENT IN SELECTION (Paper Session)

Chair: Steven S. Nettles, Assessment Systems, Inc.

Discussant: Ronald A. Ash, University of Kansas

Supplemental Application Validation Based on Self-Rating and the Suitability of Subject Matter Experts and Raters

Wilfrid N. Broderic, King County, Washington

Introduction and Problem Statement

King County conducts all types of examinations. However, the majority are variations of Primoff's checklist type of supplemental application with a heavy emphasis on background samples of actual work. One of the major concerns facing our examining program has been obtaining timely and useful validation information on an ongoing basis. Although we have over 3,500 employees, most classes are small and provide no real opportunities to conduct criterion validity studies. The feedback we receive about the results of a particular examination is often both negative and positive, containing no factual information to support the feedback and offers a few solutions. Most discussions would result in using the content validity model to defend the examining program, and this proved frustrating for the complaining parties and provided little satisfaction to candidates, hiring authorities, affirmative action representatives, union representatives, or any other concerned party.

At the conclusion of each examination administration we were unable to concretely reach any conclusions regarding the effectiveness or non-effectiveness of the job analysis, raters, rating criteria, applicant responses or any other components of the examination process. Faced with an absence of hard data in most cases, we decided to begin collecting self-ratings from the candidates and compare these to the raters' results. We felt these self-ratings could solve many problems associated with validation studies, especially timeliness and restriction of range. By using these self-ratings we hoped to achieve high and significant correlations between candidate self-ratings and raters' scores on supplemental applications. We also hoped these correlations would support our policy of using subject matter experts and raters at the level of work being examined. In addition, we hoped that if high and significant correlations were not achieved, we would at least be able to analyze and identify causes of low, and/or insignificant correlations.

Hypothesis

The first hypothesis of this paper is that a candidate's self-rating should be highly and significantly correlated with the raters' scores on supplemental applications. The second hypothesis is that people

at the level of the job being examined will make better subject experts and raters than people not at the job level.

Self-Ratings:

The self-ratings we developed are based on the same job analysis used to prepare the supplemental application. Most of our job analyses result in 4 - 6 major elements with 5 to 8 content items listed for each major element. The introductory statement of the self-ratings attempts to assure the candidates that their ratings will not be used in any way except for research. We assure them that the raters of the supplemental application will not see the candidate's self-ratings. We also tell the candidates that the self-rating data will be used to improve and check on the performance of the supplemental applications. We obtain very close to 100% completion of the self-ratings with few ratings that do not identify the strengths and weaknesses of the candidates.

The basic idea for this format was derived from Primoff's checklist questionnaire where candidates rate themselves and justify their self-rating. However, we believed that many times the raters were influenced by the self-rating in the Primoff system. In addition, exposing the raters to the candidate's self-rating and then using that self-rating as a criterion for validating the raters' evaluation was not very defensible. Our solution was to divide the Primoff checklist into a supplemental application and a self-rating questionnaire described in the previous paragraph.

The self-rating scale is a modified version of Primoff's job element scale. Initially we had some negative reactions to Primoff's job element scale and modified it until we reached maximum user comfort and use. The applicants use this scale to rate themselves on each content item listed for the major elements. At no time are the candidates informed of the points assigned to their self-ratings.

We calculate the self-rating score for each applicant by giving the following points for each content bahavior: NONE = 0; POTENTIAL = 1; ACCEPTABLE = 2; JOURNEY = 3; EXPERT = 4. We add the total for each major element and divide by the number of content items, and total the major elements' points.

Correlation Study Results:

The correlation studies for this paper were conducted in 1984 and represent almost all the supplemental application examinations which contained enough data to conduct validation studies. The data was arranged into the following three groups:

Group I -Both subject matter experts and raters are at the level of the job being examined.

Group II -Combination of subject matter experts at the job level and direct supervisors.

Group III -Subject matter experts and raters who are knowledgeable about the job but are not at that level or not direct supervisors.

While the correlations themselves will allow an examination of the validity of the first hypothesis of this paper, the groupings as previously described will allow an examination of the second hypothesis.

Analysis Of Correlation Results

Even though we stress the need to use subject matter experts and raters in the class being examined and/or their immediate supervisors, the results clearly indicate we were only able to achieve this in 19 out of the 63 cases (30%). This is due to small and/or new classes. Thus we were very interested in comparing the correlation results between groups. Even though the average correlation for Group I (.42) is higher than the average correlations for Group II (.34) and Group III (.34), the evidence is not conclusive; there are no dramatic differences between the average correlations of the three groups. These results suggest anyone knowledgeable about a job can produce a valid supplemental application.

Use Of Correlation Results

While using subject matter experts and raters from among those at the level or supervising the class being examined makes legal and professional sense, the evidence does not support the hypothesis that this is necessary to produce valid examinations. Even though these data suggest we can use other knowledgeable workers, we will continue to request subject matter experts and raters as close to the work as possible because of the positive reception by candidates and hiring authorities. In addition, in several cases we improved our correlations significantly by using people as close to the work as possible. In 1983 we used a supplemental application for Engineer - Survey which was designed by management and rated by consultants (N=20, r = .22/NS). Both the hiring authority and candidates were unhappy with the resultant list. This year (1985) we used Engineer - Surveyors as subject matter experts and raters. There were no complaints and the validity coefficient more than doubled (N=33, r = .48/.01).

The most useful function of the correlation is that it provides evidence favorably accepted by everyone to discuss problems concerning the results of the supplemental applications. As a result of the correlations and subsequent discussions, we have identified a number of problems that commonly occur. The most frequent identified ones fall into three categories:

Subject Matter Experts

- Job analysis may be done by individuals who are not totally familiar with the work as it is currently performed.

- Subject matter experts may not be superior workers, but are those who are available or are the result of other undetermined agendas.

- Subject matter experts may fail to perform job analysis duties as instructed.

- Subject matter experts may be uncooperative, have their own ideas about job analysis, or do not see any need for job analysis.

Raters

- Raters may not follow instructions.

- Raters may use their own criteria instead of what is provided.

- Raters may feel rushed or pressured and not rate as accurately as they should

- Raters may not be totally familiar with the work as it is currently performed.

Candidates

- An ineffective supplemental application format may not draw out the proper information from candidates.

We found that most of the negative or very low correlations were the result of an extreme of one or more of the above problems. Once we identify one or more of these points as the problem(s), we can take corrective action to improve the results. When we have done this, as discussed in the earlier example, (Engineer - Survey), we have had nothing but improvement in every instance.

In only one case so far, Environmental Health Specialist, have we used the same job analysis, raters, and supplemental application two years in a row. In 1983 the results were N=30, r = .61/.01, and the results in 1984 were N=20, r = .68/.01. It should be noted that 20 of the candidates on the 1983 employment list were either hired or declined an offer of employment, thus 1984 was not a repeat of the same people.

We have also correlated the self-ratings with multiple-choice examinations and obtained the same positive results as with the supplemental applications.

Conclusion:

The first hypothesis states, at the beginning of this paper, that candidate's self-ratings should be highly and significantly correlated with raters' supplemental application scores. The data support this hypothesis with an average correlation of .35 for 63 cases.

* * *

The Validity of Self-Assessments Within a Police Sergeant Promotional System

Kevin G. Love, Central Michigan University, Mt. Pleasant, MI

The few published studies of the validity of self-assessment for personnel decisions have yielded mixed results. Many of these investigations have found significant validity of self-assessments on relevant applicant characteristics (e.g., typing ability, typing speed, spelling ability, and word meaning for clerical workers) when linked with actual standardized test performance in these same areas. When an attempt has been made to link self-assessments with actual job performance, however, the few research studies have consistently shown invalidity.

The inconsistencies in the research literature do not indicate an overall invalidity of self-assessments in predicting on-the-job performance, as suggested by Hunter and Hunter (1984), merely the impact of supervisors and employees utilizing different definitions and weightings of work behaviors in describing job performance. That is, supervisor or assessor rating measures may be interpreting different employee behaviors or characteristics as indicative of good and poor job performance than intended by the organization or as seen by the employee (as revealed through their self-assessments). This common differential perspective of the job, based on organizational level and position, may moderate the findings of validity for self-assessments (Campbell, Dunnette, Lawler, and Weicke, 1970).

The current study sought to determine whether self-assessment would reveal different levels of validity (i.e., differential validity, as defined by Boehm, 1977) depending upon the nature of the criteria employed. Two previously researched types of criteria were used; traditional supervisor performance ratings; and a written test of promotability.

It was hypothesized that, using the same job performance dimensions, self-assessments would relate more highly to an examination-based measure of performance than subjective supervisor performance

evaluation ratings, based on the differential perspectives of appropriate worker characteristics between employees and their supervisors. The subjective interpretation and/or biases of the supervisors, as revealed through traditional performance ratings, would decrease their relationship with the self-assessment measures. The use of a written examination measure would minimize such biases and thus yield a higher relationship with employee self-assessments.

Method

Subjects

The present study involved 73 police officers who were under consideration for promotion to the position of police sergeant. All officers were employed within a large municipal police department and had been informed as to their promotion status.

Study Instrumentation

Job Analysis. A task-based job analysis was completed for the position of police sergeant within the subject organization. Interviews with selected sergeants were used to develop a task rating questionnaire which was completed by all sergeant personnel (n=42). Through combining ratings for each task on frequency of occurrence, level of difficulty, and consequence of error the most important tasks for the position were identified. Relevant requisites (i.e., knowledge, skills, abilities, and personal characteristics) were derived from these tasks by a committee of job analysts and police sergeants. Using a content analysis procedure the requisites were clustered into 13 performance dimensions: decisiveness, resilience, flexibility, perseverance, initiative, sensitivity, political sensitivity, impact, emotional control, self confidence, open mindedness, dependability, and leadership.

Self-Assessment Rating Instrument. Using the 13 performance dimensions developed from the job analysis data, the police sergeant candidates were required to provide ratings on each dimension as to how much they possessed each characteristic. A graphic rating scale approach was used with "1" being low and "5" being high for rating purposes.

Supervisor Rating Instrument. Supervisors of the police sergeant candidates (primarily police sergeant personnel) were required to provide ratings on the candidates control level of performance or ability within each of the 13 areas. A graphic rating scale approach was used as described above.

Written Promotional Examination. Using the 13 performance dimensions as the foundation, a 75-item written examination was developed by a team of three job analysts. Each performance dimension was

represented by a minimum of four multiple choice-type items. The written examination was reviewed by a committee of police sergeant, lieutenant, and captain personnel who provided input as to ambiguous items, potential misinterpretation, etc. The final written examination was approved by the committee and police department as the bona fide promotional examination. A pilot test of the written examination using 34 police sergeants as the sample yielded a split half reliability estimate of r = .915.

Candidate Opinion Questionnaire. In addition to the assessment instruments, each candidate was required to provide opinions regarding the use of self-assessments within the promotional process. A 4-point Likert-type scale was constructed and opinions were gathered on five reaction areas.

Seniority of Candidate. Using organizational personnel files the seniority of each police sergeant candidate was recorded as to the number of months the officer had been employed within the subject organization.

Procedure

The police sergeant candidates were required to provide the self-assessment ratings and their supervisors the performance ratings one week before completing the written promotional examination. The candidates and their supervisors were informed via a group meeting that the self-assessments, supervisor evaluation ratings, written test scores, and seniority points would be combined to provide a final determination of promotion status.

The written examination was group administered and computer scored. Opinions regarding the use of the self-assessments within the promotional process were collected one week after the written examination had been completed, but before any feedback regarding test performance had been announced.

All candidates were provided their final ranking as to the promotion list (as determined by joint police department and union agreement) at a group meeting. Individual reactions to the promotional system were also encouraged within the meeting agenda.

Results

Self-Assessment Ratings. A principle components factor analysis was used to determine whether the self-assessment ratings across the 13 performance dimensions could be reduced to a smaller set of factors. Five factors were needed to account for the variance of the self-assessment ratings: problem solving, adaptability, responsibility, leadership, emotional control. The factors were labeled based on the definitions of the performance dimensions which loaded most highly on each. The self-assessment factor scores were significantly intercorrelated. Factor scores were computed and used in subsequent analyses.

Supervisor Performance Ratings. A similar factor analysis, as described above, was performed using the supervisor performance ratings across the 13 dimensions. A significant amount of halo error within the ratings occurred as the factor analysis resulted in a single composite factor which accounted for almost all of the variance in the performance ratings. Based on this analysis an unweighted average (composite) supervisory rating across the performance areas was used for subsequent analyses.

Validity of Self-Assessment Ratings

Written Examination Criterion. Higher self-assessment factor scores for Problem Solving, Adaptability, Leadership and Emotional Control were significantly related to higher written test performance. Responsibility was not significantly related.

Composite Supervisor Performance Rating. While a significant relationship among written test performance and the composite supervisor performance rating was found, correlations between this type of criterion and the self-assessment factor scores were not significant.

Relationship of Self-Assessments and Seniority

Seniority of the candidates was not significantly related to level of self-assessment for any of the performance factors.

Candidate Reactions to the Use of Self-Assessments

The candidates were not favorable towards the use of self-assessments within the promotional process. A majority of candidates believed that the self-assessments would not be helpful in selecting sergeants, candidates would not provide honest self-assessments, the self-assessments provided little information as to what it takes to be a good sergeant, and using self-assessments would not make the competition for promotion fairer. The candidates did respond positively that the qualities measured via the self-assessment ratings should be measured within the promotional system.

Discussion

The study findings reflect not the invalidity of self-assessments in predicting performance on the job (as determined through supervisory ratings), but the impact of poor criteria within a validation paradigm, with special regard to the predictive power of self-assessments. Significant differential validity was demonstrated for self-assessments depending on whether written test score (high validity) or subjective performance ratings (zero validity) were used as criteria.

Validity Using Supervisor Rating Criteria

The findings of zero validity for self-assessments using supervisor performance ratings is consistent with past research (Farley & Mayfield, 1976; Riji & Page, 1980), yet somewhat surprising in light of the incorporation of identical performance rating dimensions for both sergeant candidates and their supervisors. Under these conditions one would expect at least a small degree of overlap in these two judgements of ability levels. Based on these data, it appears that the candidates and their supervisors had substantially different opinions of ability within each of the 13 performance areas under study. Whereas Primoff (1980) has suggested that the validity of self-assessments could be improved if there is common understanding of the dimensions to be rated among raters, the present study provided for this and found zero validity for self-assessments as predictors of supervisor performance ratings.

The supervisory ratings were shown to have poor criterion properties through the findings of a significant level of halo error (i.e., a single composite factor was revealed which accounted for almost all relevant variance in the ratings across the 13 dimensions). The finding of significant halo error in supervisory ratings is not unique to this study. Perhaps the current study also reflects the use of a "typical" rating situation for supervisor evaluations:

(a) a lack of training in providing accurate appraisals of employee performance (the supervisors in the present study had no formal training or experience with the rating instrument),

(b) ambiguity of performance dimension definitions (there may have been a significant discrepancy in the interpretation of candidates vs. their supervisors, i.e., zero correlation among candidate and supervisor judgements), and

(c) the influence of prior knowledge and/or biases regarding the candidate which are reflected within the performance ratings (all supervisors had substantial prior knowledge of candidate performance as police officers).

All of these factors contributed to the poor criterion properties of the supervisory ratings.

Validity Using Written Test Score Criteria

The study findings, however, are consistent with past research which had documented the significant relationship of self-assessments with written test performance (e.g., Ash, 1980; Levine, Flory and Ash, 1977). The self-assessment ratings revealed four performance factors which significantly correlated with a knowledge-oriented measure of police sergeant requisites.

Improving Validity of Self-Assessments

Interpretation of the study findings should not lead to a conclusion that self-assessments have little or no validity in predicting job relevant characteristics or on-the-job performance within a promotional situation. Instead, the findings show that researchers need to focus on improving the characteristics of the criteria used as indirect measures of on-the-job performance (i.e., performance ratings). One suggestion for improvement of self-assessment research is for the incorporation of behaviorally-anchored rating scales (BARS) for both self-assessments and supervisor performance ratings. The retranslation process of developing BARS (Smith & Kendall, 1963) would lessen the ambiguity among applicants and supervisors regarding performance in various rating dimensions. Heneman (1980) has suggested that a specification of job behaviors to be self-rated may improve the overall utility of self-assessments within the personnel decision process.

While many authors have provided recommendations regarding the improvement of self-assessments, the present study findings represent that the "criteria problem" plagues self-assessment research as well as more traditional selection methods. Other suggestions for improving self-assessments have ranged from improving the employee's self-esteem (Bassett & Meyer, 1968) to providing applicants with a choice as to which has become so important in the development of personnel systems. Indeed, if supervisors, peers, candidates, etc. were allowed to select only "meaningful" performance areas for their evaluation, there would be little systematic coverage of the crucial requisites needed for good job performance as identified through a careful job analysis process.

Heneman (1980) is correct, however, in the assertion that self-assessments guide external selection. The use of self-assessments within organizations, outside of selection and performance appraisal systems, is increasing in frequency (Burack, 1979). Self-assessments have been found useful in providing information for career path planning and employee development. In spite of a lack of empirical evidence documenting the accuracy of such self-appraisals, their incorporation into these decisions has been valued by employee and supervisor alike. This area needs attention as the career movement issue becomes one of systematic planning as opposed to movement based primarily on an employee's self-assessment and presumed knowledge of advancement opportunities. The use of self-assessment in the career pathing process would provide information regarding their subsequent impact on future job behaviors (i.e., promotions, lateral career moves, number of job changes, etc.).

Therefore, the process of self-assessment should not be abandoned as one without validity. The validity of this procedure is subject to the same biases and failures of rigor which plague any measuring device. The utility of self-assessments within personnel decision making has yet to be addressed from a sufficiently broad base of empirical research.

References

Ash, R. A. (1980) Self-assessments of five types of typing ability. Personnel Psychology, 33, 273-282.

Bassett, G. A. & Meyer, H. H. (1968) Performance appraisal based on self-review. Personnel Psychology, 21, 421-430.

Boehm, V. R. (1977) Differential prediction: A methodological artifact? Journal of Applied Psychology, 62, 146-152.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III & Weicke, K. E., Jr. (1970) Managerial behavior, performance, and effectiveness. New York: McGraw-Hill.

DeNisi, A. A. & Shaw, J. B. (1977) Investigation of the uses of self-reports of abilities. Journal of Applied Psychology, 62, 641-644.

Edwards, A. L. (1976) An introduction to linear regression and correlation. San Francisco, CA: Freeman & Co.

Heneman, H. G., III (1980) Self-assessment: A critical analysis. Personnel Psychology, 33, 297-300.

Hough, L. M., Keyes, M. A., & Dunnette, M. D. (1983) An evaluation of three "alternative" selection procedures. Personnel Psychology, 36, 261-275.

Hunter, J. E. & Hunter, R. F. (1984) Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Levine, E. L., Flory, A. P., & Ash, R. A. (1977) Self-assessment in personnel selection. Journal of Applied Psychology, 62, 428-435.

Primoff, E. S. (1980) The use of self-assessments in examining. Personnel Psychology, 33, 283-289.

Reilly, R. R. & Chao, G. T. (1982) Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Smith, P. C. & Kendall, L. M. (1963) Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Van Rijn, P. & Payne, S. S. (1980) Criterion-related validity research base for the D. C. firefighter selection test (PRR-80-28). Washington, D. C.: U. S. Office of Personnel Management, Office of Personnel Research and Development. (NTIS No. PB81-122087)

* * *

Departmental Ratings for Promotional Examinations

Patrick T. Maher, Personnel and Organization Development Consultants, Inc., LaPalma, California

This paper presents a model to use for departmental ratings for promotional examinations. It describes an appraisal of promotability for the position of fire captain in a large fire department. The model includes:

> Rating of specific knowledge, skills, abilities, and other characteristics (KSAs) derived from the job analysis.
>
> A systematic method to record behaviors to each of the specific KSAs and document them as a part of the annual performance evaluation.
>
> A procedure that permits each individual to protest any recorded behavior and to add additional behavior for consideration at the time that the annual performance evaluation is completed.
>
> Several options for rating (scoring) the behavior at the time that the examination is administered.

An appraisal of promotability is a viable method for determining the ability of an employee to perform at a higher level. The appraisal of promotability uses an applicant's past behavior. An appraisal of promotability, however, is not a perfect or infallible process, and like other selection procedures has certain limitations. In most instances, the biggest problem with using an appraisal promotability is lack of adequate documentation. Typically, performance evaluations do not provide sufficient behavior about dimensions being measured for promotion or are too superficial to be of much value.

In addition, when the rank for which the appraisal of promotability is being used as the first true level of supervision in the department, such as fire captain or police sergeant, it is often difficult to obtain adequate indicies of behavior on which to base the evaluation.

The candidate pool usually does not perform work behaviors that give an accurate indication of the ability to perform the types of supervisory duties required of first line supervisors.

There is, however, a method that overcomes most of these problems, provided the department and raters commit themselves to the process.

Obtaining Data

Any true appraisal of promotability system must be linked to the performance evaluation process, because both deal with many of the same behaviors during the same time frames, but a problem inherent in this is that an employee's current performance and potential performance at a higher level may be rated at two distinctly different levels. For example, an employee may be performing at an outstanding level at his current job, but may be rated unqualified when considered for promotion. It is recommended, therefore, that the appraisal of promotability form be completed annually with the performance evaluation, but that the appraisal of promotability evaluation not be considered when deciding on the annual performance rating. Further, it is recommended that no score be given on the annual appraisal of promotability material that documents relevant behavior during the evaluation period. When the employee applies for promotion, the appraisal of promotability forms from the various years can be used to determine the appraisal of promotability for the specific examination being conducted.

Employee Involvement

The importance of employee input cannot be overemphasized, although there will likely be strong opposition to it by managers, especially in paramilitary organizations such as police and fire departments. In a civil service or merit system, however, employee input is inevitable, either during the evaluation period or during subsequent civil service appeals. It is far better to identify and deal with the issues in an informal, counseling relationship than to have to deal with them much later during an adversary relationship.

Appeals

In addition, there should be an appeal process that enables any employee to appeal the accuracy of an appraisal of promotability. This appeal process can be the existing one for appealing a performance evaluation or filing a grievance, or it can be a new one designed specifically for the appraisal of promotability process.

Time

All factors considered in the appraisal of promotability process should have a time limit. Generally, any behavior or factor (whether positive or negative) more than 5 years old should be eliminated from consideration.

While a five-year cut-off period is generally recommended, independent judgement should be exercised in each case and for each position. Regardless of the period used, it should be announced, as a minimum, at the time that the examination is announced.

One problem with having a cut-off time is how to give appropriate credit for certain types of performance that the department desires to encourage and reward. The most notable example is that of education. While a person is attending school, of course, he will receive appropriate credit as he completes classes and for the first five years after completing school. With a five-year cut-off, however, the employee will eventually lose credit for education and be rated identically to the employee with no education.

Verified Data

All behavior reported on the appraisal of promotability form should be considered as verified or unverified. While both types of information are important and should be considered, the verification of as much data as possible is important to maintaining the integrity of the process. Information not verified can be easily falsified. For example, if credit is to be given for attending job-related seminars, then a candidate could claim that he attended a seminar when in fact he did not. By obtaining independent evidence of such attendance, the likelihood of false information is reduced.

Specific Behaviors

At times, there is a need or desire to consider the extent to which certain behaviors are present. For example, in this project, executives of the fire department expressed a desire to evaluate certain behaviors for the purpose of rewarding or encouraging certain conduct that they feel is beneficial to the department. These behaviors fell into one of two dimensions being assessed by the appraisal of promotability. They are listed below each of the dimensions to which they apply:

> Ability to work in a para-military organization, accepting orders and complying with established procedures without hesitation.
>
> Desire to actively influence events rather than passively accepting them; self-starting; takes action beyond what is necessarily called for, as shown in volunteering for special assignments, continuing education related to details, etc.

All factors that are to measured should be announced to the potential candidates as soon as possible so that they can adequately prepare for the examination. In the future, as other items are added, they should be announced immediately. In addition to providing the candidates with advance notice for the examination, the department will obtain some benefits when employees strive to meet the criteria that the department considers important for promotion.

Ratings

Ratings themselves can be made at the time of the examination process through several different methods. The department can choose the method that best suits its desires and needs, but should do so well in advance of the examination. The rating procedure should be stated in writing so that everyone is made aware of it.

Any acceptable procedures involves the candidate's immediate superior preparing a composite narrative based on the appraisal of promotabiltiy documentation in the personnel file. In this process, the rater merely consolidates the information from the different years into one comprehensive description within each area being rated.

The candidate has an opportunity to review the summary and append to it a statement that includes his comments concerning accomplishments and other factors that he feels impact his qualifications.

Ratings are assigned through a group consensus with a panel consisting of a minimum of 3 supervisors. More can be included if desired. When ratings are made, they can be done with either the identity of the candidates known or not known. In the first ones, which is usually preferred by most raters, raters have an opportunity to add their own perspective on the candidate if they feel that it is not adequately covered in the documentation. If the documentation process is not comprehensive, then such input can be critical to a proper system. On the other hand, raters tend to substitute their subjective opinions (especially bias) for the objective data, undermining the purpose of collecting the objective data. This is especially true where the identity is known to the rater.

In the second case, raters are provided only with the narrative data concerning the candidates. Names are omitted from the narrative portion, and a number is assigned to each candidate. This process forces the raters to make their judgements entirely on the basis of documented data, thereby eliminating many rating errors, in particular bias and halo. One problem with this is that some raters may recognize individuals by some of the documented material, while not recognizing others. Thus, an imbalance is created in the level of subjectivity versus objectivity. Another problem is that a rater may have specific facts that are relevant to the rating of an individual, but not knowing who he is, cannot consider such facts. Again this becomes especially critical where documentation is lacking. Whichever process is used, all ratings are made by the panel on a 5-point scale. Definitions of the rating scale levels should be provided.

Summary

In summary, the use of an appraisal of promotabiltiy is a viable process that provides some of the best information on future performance, provided that the appraisal of promotability process is properly designed and administered. In adopting this process,

the department will have to make certain decisions, publish them, and implement them or take other action. Decisions or actions to be made or taken by the department, as indicated above, are summarized below:

Development and implementation of a training program for raters.

Implementation of a documentation process, including appeals, input from the candidates, and verification of information included.

Establishment of a time period for factors to be considered, including any exceptions to the time period.

Establishment of a rating process that includes who will serve as raters, and what information they will have (e.g., identity of candidates, etc.).

Publication of all decisions and processes, and updating of such decisions and processes as they are modified.

References

Epstein, S. & Laymon, R.S. Guidelines for Police Performance Appraisal, Promotion and Placement Procedures. Washington, D.C.: U.S. Department of Commerce, National Technical Information Service, March, 1973.

Schmidt, F.L., Caplan, J.R., Bemis, S.E., Decuir, D., Dunn, L. & Antone, L. The Behavioral Consistency Method of Unassembled Examining. (TM-79-21) Washington, D.C. U.S. Office of Personnel Management, Personnel Research and Development Center, November 1979.

Zedeck, Sheldon. Performance Measures: Forms or Samples? Summary of an invited talk at the Annual Conference of the International Personnel Management Association Council, May 1984, Seattle, Washington.

* * *

THE USE OF EMPLOYMENT SELECTION PROCEDURES WITH LARGE MULTI-ETHNIC AND RACIAL CANDIDATE POPULATIONS: PERSPECTIVES AND STRATEGIES (Symposium)

Chair: Priscilla J. Hambrick-Dixon, New York City Department of Personnel

A Methodology to Determine Job Required Reading Levels

Anthony Roig, and Priscilla J. Hambrick-Dixon, New York City Department of Personnel

Introduction

The ability to satisfactorily perform in many jobs is determined to some extent by the individual's ability to read. To enhance the job-relatedness of employee selection procedures requiring reading competency, it seems beneficial to assure that the selection procedures are as closely linked as possible to actual job reading requirements (Stricht, 1975).

Section 14 of the Uniform Guidelines on Employee Selection Procedures stresses the importance of a selection procedure closely approximating the work situation when establishing content validity. Although the Guidelines are explicit in providing technical standards for establishing content validity, they provide no specific methodology to evaluate this correspondence.

The intent of this paper is to provide professionals involved in test construction a means by which the similarity between the reading level of the test and the level of reading required on the job can be evaluated. It is thought that matching the reading level of the test with the reading level required on the job would provide support for content validity.

Computerized Readability Analysis Methodology

The methodology advocated in this paper involved a series of steps focusing on the use of the readability program commercially available on a disk. The Readability Program requires an Apple II computer with 48K Applesoft in ROM and a single disk drive. The Readability Program measures the level of text material according to nine different formulas: Dale-Chall, Fry, Flesch-Kincaid, Fog, ARI, Colman, Powers and Holmquist. The data analyzed were collected during a job analysis conducted by the New York City Department of Personnel in its involvement with an unskilled, entry-level Civil Service examination. Since New York City has a multi-ethnic and multi-racial candidate population, the Department of Personnel staff was concerned about the possibility of adverse impact.

Similar to Payne's (1983) methodology, the first step was to obtain an objective measure of the reading level required on the job. A variety of reading materials such as manuals, teletypes, orders, etc. used by incumbents were collected and analyzed by the Readability Program. The Readability Program provided reading levels based on the nine different readability formulas mentioned earlier. However, data obtained from the job analysis indicated that incumbents received job training predominantly from oral instructions. This information clouded the accuracy of the reading level needed on the job.

The second step consisted of verifying the accuracy of the reading level needed on the job. Staff involved with the training of the job incumbents were summoned and asked to identify materials incumbents must read before receiving training. These materials were then analyzed by the Readability Program and reading levels were obtained as well. These reading levels were compared with the previously determined reading levels, a step recommended by court proceedings (Payne, 1983).

The third step was to investigate the list of difficult words which were identified by the Readability Program. Many words identified were job jargon words used by incumbents (i.e., refuse, commercial establishment). According to EEOC Guidelines, knowledge of these words should not be tested for in an entry-level open competitive examination. Since this was such an examination, Department of Personnel staff reasoned that these words can be easily changed, without affecting the validity of the test and to avoid the possibility of adverse impact. Further, it was thought the use of these words may appear as discriminatory to those candidates taking the test who are not familiar with the job jargon terminology. Therefore, Department of Personnel staff decided that these words would be replaced with similar but neutral words on the examination.

The final step was to determine the reading level of the test questions and match them with the reading level obtained from the required reading material. In all cases, the test questions were found to be at or below the reading level required on the job.

Discussion

While there are many valuable uses of the Readability Program, there are some apparent problems which accompany its use as well. These are as follows:

- mistakes, when entering text data, are not easily corrected;
- the program does not directly support the variety of printers available for use with the Apple computer;
- the program also does not indicate which formula identifies a word as difficult.

In addition, readability formulas rarely agree as they are not designed to measure the same reading level. However, one trail

common with all formulas is the necessity of manually counting components of text to apply against the rules of the various readability formulas. This process is both time consuming and subject to error. By using the Readability Program, several formulas can be calculated at once with no corresponding increase in effort and a trend can be identified.

In the final analysis, it is the user who must determine which formula is the appropriate one to use in the given situation. The Readability Program allows the user to concentrate on this role rather than the time consuming role of counting and mathematical calculation which is necessary before true diagnosis can begin.

It is thought that such methodology would assist test constructors in determining the reading level required on the job and in constructing the most valid and defensible test possible.

References

Readability Program. Opportunities for Learning, Inc. 8950 Lurline Ave., Dept. W, Chattsworth, California 91311

Payne, Sandra. Problems in Job-Related Measurement of Reading Ability. Paper presented at the meeting of the International Personnel Management Association Assessment Council, Washington, D.C., May 1983.

Sticht, Thomas G. Reading for Working: A Functional Literacy Anthology, Human Resource Research Organization, Alexandria; Virginia, 1975.

* * *

A Systematic Approach to Determining Critical Job Behaviors

Jorge L. Esquilin, New York City Department of Personnel; Eric S. Stein, New York City Department of Personnel and The College of Human Services; Miguel Roig, New York City Department of Personnel and Montclair State College; Ewald Weber, New York City Department of Personnel.

Sections 14 and 15 of the Uniform Guidelines on Employee Selection Procedures address the standards for validity studies and the documentation necessary when such studies are undertaken. These sections are explicit in terms of criteria ("what" and "when"). However, the Guidelines do not focus upon a systematic approach ("how") to this problem.

Numerous public agencies, as well as private corporations, engaged in personnel testing have interpreted and applied the Uniform Guidelines. However, their interpretations are often fragmented, provincial and procedurely unsystematic. The intent of this paper is to present and advocate a sound and psychometrically acceptable methodology to assist test constructors in determining the critical and/or essential job behaviors.

The methodology presented in this paper was developed from the data collected during two job conducted by the New York City Department of Personnel, in its involvement with Civil Service promotional examinations. As suggested by the Uniform Guidelines, job behaviors were rated on five Likert-type scales: Frequency, Importance, Level of Difficulty, Time Spent and Consequences of Error. Pearson Product Moment correlation coefficients indicated that scores on the "Frequency" scale were related with those on the "Time Spent" scale. Similarly, "Importance" was highly correlated with "Consequences of Error" suggesting the collapse of scale scores into two new variables. Moreover, multiple regression analyses produced significant findings impining upon the determination and documentation of content validity as called for in the Uniform Guidelines, and more recently in many court cases.

(Note: No data were made available by the authors, so only an abstract could be presented.)

* * *

Strategies and Outcomes of Developing a Written Examination for a Large Multi-Ethnic and Multi-Racial Candidate Population

Charles S. Wachter, and Priscilla J. Hambrick-Dixon, New York City Department of Personnel

For many years, psychologists and educators have debated the cause of and conducted much research on the phenomenon of differential performance of minorities (particularly Blacks and Hispanics) on written examinations. The prevailing theories for differential performance of Whites and minorities on tests include nature vs. nurture; that is, that intelligence is determined by biological or environmental variables or by a combination of both.

From an environmental theoretical perspective, Olmedo (1981) espoused that psychological and educational testing of members of minority groups should take into account the diverse social, political, and economic realities currently facing Blacks and Hispanics. Referring to these groups as "linguistic minorities", he believed that linguistic issues related to bilingualism and acculturation had many implications for testing of multi-ethnic and multi-racial groups. Thus, important considerations for the assessment of "linguistic minorities" using a written mode (i.e., reading comprehension aspects of tests) are "what performance is required on a given written test?"

For many decades, the issue of readability of written material has been investigated as it relates to test construction. The term "readability" in employment testing refers to an indication of the ease of understanding or comprehension due to the type of writing or content (Klare, 1963). One of the earliest renowned applications of readability analysis in the employment testing arena was by Rudolf Flesch (1974).

With regard to test development, many researchers posit that the conduct of readability analysis in the test development process has three major benefits. It 1) increases validity; 2) aids affirmative action and 3) promotes good public relations (Allen).

Payne (1983) developed a methodology for conducting a readability analysis to construct a job-related reading comprehension test for firefighters. This methodology involved the following steps: 1) A task-based job analysis was conducted to identify all of the critical abilities required for entry-level job performance (a completed task inventory was used to determine the important tasks of the job); 2) A panel of firefighters identified the abilities that were required to perform these important tasks and that should be included in a written examination to be used to rank job applicants. (Reading comprehension was one of the abilities identified); 3) Written materials used in entry-level job performance were identified by the panel of firefighters; 4) A Flesch reading ease index was computed for each of the required written materials to get an objective measure of the actual job reading level. The index was also used during test development to ensure that the reading comprehension part of the written test was set at a level actually required by the job; 5) A multiple-choice test item reading comprehension test was designed with paragraphs taken from actual firefighters job materials. Each question could be answered by a careful reading of the paragraph. No special experience or training would be required or would help in choosing the correct answers; 6) All of the reading comprehension test items were pretested on groups of applicants for clerical positions in the Federal government.

It was evident that the development and outcomes of the test developed by Payne (1983) had many implications for the development of a proposed qualifying written test for New York City's Sanitation Workers (garbage collectors). This test was to be administered to a multi-ethnic and multi-racial candidate population (approx. 82,000). Additionally, there was a legel mandate pending from the previous test -- by Hispanics -- that the next test to be administered to Sanitation Workers was to be reviewed by language consultants. A court injunction was very likely if this mandate had not been observed.

The New York City Department of Personnel's major concerns in the development of the qualifying written test were that the test be job-related, set at the correct reading level and contain as little ambiguity as possible. Our task differed somewhat from Payne's in

that the Sanitation Worker position involves mainly laborer-type activities with minimal need to refer to reading material as an integral part of the performance of job tasks. Thus, our key consideration was that the entry-level reading comprehension test correctly concentrate on the typical required daily reading tasks and not on formal job prescriptions.

In light of the above, the following steps were added to Payne's (1983) methodology:

1) All reading source material (administrative procedures and orders, training materials and manuals, etc.) - were scrutinized in terms of frequency of use by incumbents themselves and the availability or location of the reading materials. In our initial meetings with high-level job knowledge experts, we detected a tendency on their part to report their perception of the requirements of the whole job rather than actual performance required of newly-appointed workers. This situation led us to employ our investigatory skills in an attempt to determine the actual reading that low seniority Sanitation Workers must do in performance of their daily tasks. In their case, we discovered from first line supervisors and Sanitation Workers that it was inappropriate to excerpt high school level reading material from equipment maintenance manuals, since only a small percentage of new workers must read these manuals. Instead, we were pointed to everyday teletype orders and memos and equipment and vehicle operation manuals that all Sanitation Workers must read and comprehend.

2) Two language consultants (one Black and one Hispanic) were employed to review the test and identify particular words, phrases or idioms which may be confusing or have different meanings for Hispanics and Blacks. For example, "assist" in English could be confused with "assistir a" (to attend) in Spanish and "confirm" might be confused with "conformarse" (to be satisfied with) in Spanish (de Martinez, 1979). Likewise, to refer to the "cab of a truck" or "teritiary streets" would probably be unfamiliar and/or confusing terminology to minorities. As a result of this review, the reading level of the test was lowered somewhat. We believe this a more defensible procedure than arbitrarily reducing the reading level to avoid or reduce adverse impact.

3) A readability analysis, using the Flesch test, was conducted to determine the reading level of the materials cited in (1) above which are read <u>frequently</u> by incumbents. The reading level ranged from the sixth to the eighth grade. This might be termed the root or core reading level required to perform the job at entry.

The results of the test item analysis indicated that the test was relatively easy, yet had good discrimination and high internal consistency and reliability. Most importantly, there was no adverse impact of the test on minorities and no subsequent lawsuits. Thus, the process cited here should help ensure that those who are eliminated from the selection procedure because they do not possess the language proficiency required by the test items (O'Brien, 1985) are unlikely to later show that the test did not meet <u>Uniform Guidelines</u> for job-relatedness.

Reference Notes

Allen, Peter. Language Guide for Test Construction.

O'Brien, Michael L. Psychometric Issues Relevant to Selecting Items and Assembling Parallel Forms of Language-Proficiency Instruments.

Payne, Sandra. Problems in Job-Related Measurement of Reading Ability. Paper presented at the meeting of the International Personnel Management Association Assessment Council, Washington, D.C., May, 1983.

References

Flesch, Rudolf. The Art of Readable Writing. New York: Harper & Row, 1974.

Klare, George. The Measurement of Readability. Iowa: Iowa University Press, 1963.

Moreno de Martinez, Marta. Confusing Words in English and Spanish. Puerto Rico: Editorial Mester, 1979.

Olmedo, Estaban L. Testing Linguistic Minorities. American Psychologist, 1981, 36, 1078-1085.

* * *

DEFENDING SELECTION DECISIONS (Paper Session)

Chair: Wendy Steinberg, New York State Department of Civil Service

Discussant: Paul Thomas, State of Alabama

Union Challenges to the Use and Interpretation of Promotion Examinations

Daniel G. Gallegher, Graduate School of Public and International Affairs, University of Pittsburgh; and, Peter A. Veglahn, College of Business Administration, James Madison University, Harrisonburg, VA

The purpose of this presentation and supporting paper is twofold. Attention is directed to the examination and discussion of the grounds upon which unions seek to contractually challenge the promotion decisions which are based in part on examination results. Also examined are the criteria which have been utilized by arbitrators in assessing union and employer contentions concerning the use, interpretation, and application of promotion examinations. The analysis is based on a study of seventy-six (76) public and private sector arbitration cases which involve varied union challenges to the use of promotion examinations on the basis of contractual seniority and ability language, fairness, job relevance, and the importance of nontest related criteria.

Public sector unionism has grown dramatically in the past two decades. In contrast to union intervention in the private sector, the issue of union challenges to promotion examinations within the private sector is more complex due to a number of legal and historical issues. Unlike their private sector counterparts, unionized employees within the public sector fall under the jurisdiction of hundreds of bargaining statutes, ordinances and/or executive orders. This "patchwork quilt" approach to the "law" of collective bargaining in the public sector results in differing legal definitions of the scope or subject matter of union-management negotiations. With regard to the specific focus of this paper--promotion examinations--a number of governmental jurisdictions either specifically exclude or explicitly fail to include issues relating to employee examination and promotion from the scope of contract negotiations. Within such jurisdictions, the opportunity of a union to address promotion decisions within the collective bargaining agreement may be eliminated or seriously constrained. At the other end of the spectrum, a number of government jurisdictions do allow for a scope of bargaining which is more consistent with the National Labor Relations Act (NLRA). In such jurisdictions, the opportunity exists for a public sector union to negotiate and establish parameters around promotion decisions (i.e., the relative role of seniority or promotion procedures).[2]

However, the issue of union involvement (contractual) in promotion examinations and decisions is further complicated by a second legal dimension which addresses the interface between bargaining and merit systems. Three broad legislative and/or judicial approaches can be identified: 1) the exclusion of merit system matters from the scope of collective bargaining; 2) allowing unions to negotiate some or all matters which are under the scope of civil service or merit systems--provided the contract does not conflict with the "intent" of the merit principle; and 3) to allow bargaining agreements to supercede merit-civil service rules and regulations. These distinctions both with regard to the scope of bargaining and its interface with existing merit systems is important for the reason that, unlike private sector unionism, the ability of unions to contractually address and grieve decisions resulting from management's use of promotion examinations may differ substantially from one government jurisdiction to another.

From a historical perspective, it can also be suggested that in contrast to the private sector, the use of promotion examinations has been more widespread and ingrained in the public sector employment environment. The issue as to whether or not such experience has led to greater union acceptability of promotion testing or if such a history has resulted in the aforementioned conditions of legislative and judicial protection is less clear. However, there is sufficient evidence to suggest that when given the opportunity, public sector unions would prefer to place contractual limitations on managerial discretion in promotion decisions.

Summary

Our analysis of arbitration cases involving the use of testing for promotion qualifications sought to identify the underlying factors which contribute to successful union challenges and management defenses. The results are somewhat ambiguous at this stage. The ambiguity is, to a large degree, attributable to the fact that the diversity in jobs, contractual clauses, personal and work characteristics of the grievants, and the experience of the particular union-management relationship make the process of comparative analysis and standardization excessively difficult. However, our reading and analysis of both pre- and post-Griggs challenges to promotion testing and arbitrator review have led to the formulation of some general impressions.

Most salient is our observation that the increased concern and visability of test construction and validity which have surfaced since Griggs v. Duke Power have not appreciably spilled over into the arbitration tests to measure ability. Unions have mostly continued to rely upon simply surface arguments of unfairness. It is important to note that in cases where union representatives have illustrated greater competence and ability to raise challenges on the basis of test validity and test construction, the unions have tended to have their grievances more often sustained. Despite this observation, it is also our impression that arbitrators have not generally evidenced a change in level of expertise in the assessment of management decisions which are based on testing results. The review of more recent arbitration cases suggests that arbitrators still rely heavily upon the four (4) previously noted Elkouri and Elkouri standards which have been the mainstay of arbitrator decision making in testing cases for the past four decades. Although such standards appear to provide a proper basis upon which to examine testing issues (especially job relatedness), the judgement of arbitrators tends to rely heavily upon most basic face or content validity determinations. The general absence of evidence of increased arbitrator scrutiny may be reflective of the fact that the challenges and evidence provided by the parties often focuses more upon contractual language interpretation than testing standards. Under such conditions, many arbitrators share a general reluctance to raise and pursue issues of relevance which are not raised by the parties themselves.

Our general and impressionistic conclusion is that the post-Griggs era has not evidenced more stringent and informal challenges to the use and evaluations of ability testing. However, there are a number of notable exceptions. In particular, there appears to be some suggestion that when the promotion grievance is tied to a potential Title VII complaint, the evidence and sophistication found in both union challenges and management defenses appears more informed. This observation is also extended to the quality and depth of the arbitrator's review and decision. A possible

explanation for such a finding may rest in the fact that the expectation of having to address the case in a more formal and critical judicial environment may encourage the parties to be more extensive and deliberate in the preparation of their cases. From the arbitrator's position, the knowledge that his/her decision may be raised in a second forum may produce a greater concern about the applicability of external law and testing standards.

Despite of the success which employers have appeared to enjoy in union challenges to the use and interpretation of promotion tests, we feel that this should not necessarily serve as the basis for future employer optimism. In one regard, the right of management to utilize tests to assess employee qualifications for promotion and balance such qualifications against employee seniority is fairly well established. Our general impression is that management's success in promotion testing cases has, in part, been a result of the lack of union sophistication in the area of testing. It is our observation that under more formidable challenges many of the examined management interpretations and uses of promotion testing results would be overruled in the union's favor. As an illustration, we found a number of questionable measurement practices which involved the use of cutoff scores and employee test result rankings which could be more seriously challenged by unions to their advantage.

In summary, we find that during the past decade only modest changes have occurred in the arbitration of promotion testing disputes. In comparison to more formal judicial and governmental agency review, the challenges and defenses presented within the forum of arbitration tend to be substantially less rigorous. Such a difference may, in fact, be more reflective of the differences between the contractual and legal basis of the complaint.

Footnotes

1. Within this study the term "promotion examinations" also refers to the use of examinations for job transfers and entry into training programs.

2. For an analysis of promotion grievances in the federal sector, see F.D. Ferris (1979).

* * *

References

Ace, M.E. (1971). Psychological testing: Unfair discrimination? Industrial Relations, 10, (3), 301-315.

Biddle, R.E. and Jacobs, L.M. (1968). Under what circumstances can a unionized company use testing for promotion? Personnel Psychology, 21, (2), 149-177.

Cooper, G. and Sobol, R.B. (1969). Seniority and testing under fair employment laws: A general approach to objective criteria of hiring and promotion. Harvard Law Review, 82, 1598-1679.

Elkouri, F. and Elkouri E. (1973). How arbitration works. 3rd edition, Washington, D.C.: Bureau of National Affairs.

Ferris, F.D. (1979). Remedies in federal sector promotion grievances. The Arbitration Journal, 34, (2), 37-43.

Fossum, J.A. (1977). Multiple dilemmas in testing: Professional standards, Griggs requirements, and the duty to bargain. Labor Law Journal , 28, (2), 102-108.

Hagglund, G. (1969). Psychological tests and grievance arbritation, in G. Hagglund and D. Thompson, eds., Psychological Testing and Industrial Relations! The University of Iowa (monograph series #14), 18-31.

Howard, W.(1957). The role of the arbitrator in the determination of ability. The Arbitration Journal, 12, 14-27.

Healy, J. (1955). The ability factor in labor relations. The Arbitration Journal, 10, 3-13.

Howard, W. (1958). The criteria of ability. The Arbitration Journal, 13, (4), 179-196.

Howard, J. (1959). Interpretation of ability by arbitrators. The Arbitration Journal, 14, 117-132

Jacobs, L.M. and Biddle, R.E. (1967). A review of arbitration cases involving tests for promotion. Research report, Lockheed-California Company.

McDermott, T. (1970). Types of seniority provisions and the measurement of ability. The Arbitration Journal, 25, (2), 101-124.

McConkey, D. (1960). Ability v. seniority in promotion and layoffs. Personnel, 37, (3), 51-57.

Metzler, J.H. and Kohrs, E.V. (1964). Tests and a marked difference in ability. The Arbitration Journal, 19, 229-235.

Newell, R. (1968). Psychological testing and collective bargaining. American Federationist, Washington, D.C.: AFL-CIO, February.

Slichter, S., Healy, J.J. and Livernash, E.R. (1960). The impact of collective bargaining on management. Washington, D.C.: Brookings Institution, (see 178-210).* * *

Resolving Affirmative Action and Assessment Conflicts: One Jurisdiction's Journey Through the Realm of the Possible

Samuel J. Bresler, Computer Sciences Corporation, El Segunda, CA

This paper presents a number of lessons learned from the struggles of one jurisdiction that chose neither to ignore the impact of its entry-level and promotional assessment processes on females and minorities, nor to cast aside the substantial performance benefits to be gained through the continued use of job-related selection procedures. This same jurisdiction chose to defend itself simultaneously against three legal challenges to its hiring and promotion practices, challenges initiated by organizations representing the widest possible array of social values and beliefs. Our "journey through the realm of the possible" takes us to Washington, D.C... to a discussion of the recent District of Columbia Firefighters case.

The District of Columbia Government issued an Affirmative Action Plan to govern the hiring of entry-level firefighters and the promotion of candidates to officer ranks: to Sergeant, Lieutenant and Captain positions within the Fire Department. Let me begin by describing where the District is at the present time with regard to its Fire Department hiring and promotion practices. As of today, June 20, 1985, affirmative action at the entry level is very much alive and well within the Fire Department. The entry-level provisions of the Affirmative Action Plan were modified in order to include a discussion of certification strategies that were considered and rejected by the District Government. With this brief addition, these provisions were resubmitted as a separate document to the United States District Court for the District of Columbia and approved by Judge Charles Richey. The District of Columbia Government has moved ahead in the processing of firefighter candidates, and fully expects to have a training class formed within the next few weeks. I should add that the Justice Department has appealed to the United States Court of Appeals for the District of Columbia Judge Richey's decision to approve the entry-level provisions of the plan. To date, however, the appellate process has not prevented the District Government from processing and hiring entry-level firefighters.

With regard to the promotional provisions of the plan, the District has modified and expanded the range of factors to be considered by the Fire Department Board in determining the names of candidates to be referred for promotion. Judge Richey is currently reviewing these modified provisions. His decision should be announced within the next few weeks. Promotions to the ranks of Sergeant, Lieutenant and Captain, which have been frozen during the past ten months, will not resume until Judge Richey's approval of the plan has been received. Now let me discuss the historical issues surrounding

the development of the Affirmative Action Plan. The Affirmative Action Plan was issued in accordance with the results of a four year process of administrative challenge regarding the Fire Department's employment standards and practices, culminating in a lawsuit brought to enforce compliance with the results of this administrative process. The principal administrative challenges, and the lawsuits, were supported by the Progressive Fire Fighters Association of Washington, D.C., an advocacy group of minority firefighting professionals within the District of Columbia Fire Department. Rather than marching through four years of administrative complexity, I would like to focus on the results of the administrative process by sharing with you the Hearing Examiner's findings of fact, legal conclusions and recommendations. The report submitted by the Hearing Examiner found that the District's entry-level firefighter examination, a cognitive abilities instrument, had an adverse impact on blacks when used as a ranking device, although it was noted that the entry-level examination did not have an adverse impact when used solely as a pass-fail instrument. It was also determined that the examination had not been validated in accordance with the Uniform Guidelines on Employee Selection Procedures. Finally, the Hearing Examiner found the Fire Department's promotional examinations to be job-related, content-valid processes. The Examiner recommended that the entry-level examination be validated in accordance with the Uniform Guidelines; that the District exhaust the list of all candidates who passed the 1980 administration of the entry-level examination; that all persons appointed from the 1980 list receive the same date hire, regardless of their actual date of employment, and that the Fire Department adopt and implement an Affirmative Action Plan. These recommendations were adopted virtually without change by the District Government in November of 1983. To validate the entry-level examination, the District Government selected the predictive, criterion-related model. This approach necessitated the readministration of the examination in order to assure the availability of a sufficiently large sample of subjects for the study. Before the examination could be readministered in the Spring of 1984, the Progressive Firefighters filed a suit to compel the District to comply with the results of its own administrative process. From March until May, 1984, some difficulty negotiating sessions took place, during which it was explained that before the District could complete the validation of the entry-level examination, it would be necessary to readminister the test. The Progressives quite legitimately indicated that the District had not yet published an Affirmative Action Plan for the Fire Department. On May 23, 1984, the District and the Progressive Firefighters entered into a consent decree, in which the District was permitted to readminister the entry-level examination, in which the District once again pledged to validate the entry-level examination in accordance with the Uniform Guidelines, and in which the District agreed to submit to the Court a proposed Fire Department Affirmative Action Plan. I should add that the decree made clear that it was neither an admission nor a finding of fact that the District Government had violated any law or regulation regarding prohibited discrimination.

On February 7, 1985, the Fire Department submitted to the Court its Affirmative Action Plan. On March 8, 1985, the Fire Chief promoted five black firefighters to the rank of sergeant. These firefighters represented the five highest-scoring blacks who had not yet been promoted from the October 16, 1982 Register of Eligible Candidates. These promotions were made retroactive to October 15, 1984, the final effective day of the two year register. Because there were several higher-ranking whites, these black firefighters would not have been promoted to the rank of Sergeant for reasons other than that of race. That same day, eight white incumbent firefighters and Local 36 of the International Association of Firefighters filed a complaint challenging only the promotional provisions of the Affirmative Action Plan. Incidentally, Local 36 is the sole recognized collective bargaining agent for District of Columbia Firefighters, Sergeants, Lieutenants and Captains. The complaint, which was based on Title VII of the Civil Rights Act of 1964 and the Fifth Amendment to the Constitution, alleged that the promotional provisions of the plan were illegal and unconstitutional because they contained racial preferences. On March 11, 1985, the Federal Government, through the Attorney General, filed a complaint against the District charging a pattern or practice of discrimination, again in violation of Title VII and the Fifth Amendment because the promotion and entry-level hiring provisions of the plan required preferences based on race, color or sex.

As a final note, I should mention that the original complainants, including the Progressive Firefighters, remained very much active in their proceedings against the District Government, effectively claiming that the Affirmative Action Plan did not go far enough in assuring the application of affirmative action objectives within the Fire Department. Thus, in March, 1985, the District found itself in the unenviable position of defending itself against suits from the Justice Department, the International Association of Firefighters and the Progressive Firefighters. To consolidate these cases, and to expedite the issuance of a decision on the legality and constitutionality of the Affirmative Action Plan, Judge Richey permitted all of the parties, including the District Government, to file cross-motions for summary judgement. An extensive review of the positions of each of the parties was provided through an extended hearing held on March 23, 1985. At this time, the NAACP Legal Defense and Educational Fund and the Lawyers Committee for Civil Rights each filed an Amicus Curiae brief. Judge Richey issued his ruling on April 1, 1985.

Before we discuss Judge Richey's ruling, I would like to share with you the substance of the Affirmative Action Plan—to advise you of its fundamental requirements. As a preliminary step in this process, let me convey to you the District Government's statutory authority for affirmative action, as well as certain basic statistical information about the race, ethnicity and sex composition of the Fire Department.

Statutory authority for the issuance of the Fire Department's Affirmative Action Plan derives from D.C. Code Section 1-508, which states that "Every District government agency shall develop and submit to the Mayor and Council an affirmative action plan." Section 1-507 further states that "the goad of affirmative action in employment throughout the District Government is, and must continue to be, full representation, in jobs at all salary and wage levels and scales, in accordance with the representation of all groups in the available work force of the District of Columbia, including, but not limited to, Blacks, Whites, Spanish-Speaking Americans, Native Americans, Asian Americans, females and males." Available work force is defined as the total population of the District of Columbia between the ages of 18 and 65. This became the long range goal of the Fire Department affirmative action plan during its duration, a two-year period extending from October 1, 1984 to October 15, 1986. To provide you with a brief description of the sex, race, and ethnicity percentages, the long range goal at all levels including Firefighter, Sergeant, Lieutenant and Captain, was established as follows: Blacks - 64.1%; Whites - 31.0%; Hispanics - 3.3%; Asian Americans - 1.3%; and Native Americans - 0.2%. Each of these percentages was partitioned approximately equally between males and females.

Now, what was the position of the Fire Department on April 1, 1984, the date that was used for statistical reporting pruposes, with regard to the achievement of these long range goals? First, for the entire Fire Department:

Blacks - 38.0% (37.0% black males, 1.0% black females)
Whites - 61.9% (61.8% white males, 0.1% white females)
Hispanics - 0.1% (0.1% Hispanic males, 0.0% Hispanic females)
Asians - 0.0%
Native Americans - 0.2% (0.1% native american males, 0.1% native american females)

For the officer ranks, no females or race/ethnic groups other than blacks and whites are represented. The percentages for these groups are as follows:

Sergeants	Lieutenants	Captains
Blacks - 31.6%	Blacks - 29.2%	Blacks - 15.6%
Whites - 68.4%	Whites - 70.8%	Whites - 84.4%

We can readily determine that there were gross disparities between the proportion of minorities in the available workforce and their representation in the officer ranks in the District of Columbia Fire Department.

We may now turn to a discussion of the content of the Fire Department Affirmative Action Plan. Let us begin with a presentation of the entry-level provisions. As I mentioned previously, the entry-level examination was re-administered in 1984, to a candidate group of 1626 individuals who met the Department's age, education/experience

and citizenship qualifications requirements. This group consisted of 1050 blacks (64.6% of the total group), 492 whites (30.3%), 37 hispanics (2.3%) and 47 members of the other racial groups or of unspecified race (2.8%). The group also consisted of 6.8% females and 93.2% males. First, the Affirmative Action Plan called for the setting of a passing score at a level that met the Uniform Guidelines' 80% benchmark for determining adverse impact. To ensure that no protected candidate group was disproportionately excluded from subsequent employment consideration, 1384 individuals were permitted to pass the examination. This group was composed of 830 blacks (60.0% of the total group passing the examination), 486 whites (35.1%), 33 hispanics (2.4%) and 35 members of other racial groups or of unspecified race (2.5%). This group also consisted of 96 females (6.9%) and 1287 males (93.1%).

There are, of course, two different types of adverse impact. Having dealt with the pass-fail issue, the District looked carefully at the manner in which minorities and females distributed themselves throughout the entire range of passing scores. It was determined that whites, and particularly white males, clustered near the top of a score-ordered listing. For example, in reviewing the scores of the top one hundred candidates, it was observed that 79 were white males. It was clear that the selection of candidates strictly on the basis of rank order would result in a severe adverse impact on minorities and females. The District's response was to design a procedure to eliminate such results.

The plan called for the establishment of 12 certificates or lists of eligibles, each of which consisted of approximately 120 candidates. These certificates were created after the District first generated separate lists of white males, white females, black males, black females, hispanic males, hispanic females and other males based on the candidates' scores on the written examination (plus veterans' preference points). The plan directs that the race of sex composition of each certificate approximate the pass rate for blacks, whites, hispanics, others, males and females. That is, that each certificate be composed of 60% blacks, 35.1% whites, 2.4% hispanics, 2.6% others, 93% males and 7% females. For example, the first certificate of 121 individuals to be generated would consist of the highest scoring blacks, whites, hispanics, others, males and females in sufficient number to assure their appropriate proportionate representation. This would mean that the first certificate would consist of 66 black males, 7 black females, 40 white males, 2 white females, 2 hispanic males, 1 hispanic female, 2 other males and 1 other female.

Finally, according to the Affirmative Action Plan, candidates must be monitored through the remaining stages of the Firefighter selection process (a medical examination and background investigation) and, after completion of these stages, offered positions as Firefighters in numbers sufficient to assure that each fire training class would be at least 60% minority and 5% female.

Before turning to Judge Richey's reaction to these provisions of the Affirmative Action Plan, I would like to discuss briefly the Plan's promotional provisions. For each of the officer ranks, the Affirmative Action Plan addressed the issue of disparities between whites and minorities by setting short range promotions to be made during the two-year life of the 1984 promotions register, as follows:

> For the rank of Sergeant, out of 45 promotions, 14 minorities including one female (representing 29% of the promotions to be made); for the rank of Lieutenant, out of 33 promotions, 10 minorities (30%); for the rank of Captain, out of 27 promotions, 17 minorities (63%).

To achieve these goals, the Plan requires that promotions be made on the basis of the "Rule of Nine Plus." This rule states that for each position or set of promotional positions to be filled, nine additional names will be submitted to the Fire Department's Promotion Board, in order of their promotional examination scores. For example, if there are three vacancies for Sergeants, the highest scoring three plus nine, or twelve, names will be submitted to the Promotion Board. The Board then refers to the Chief only the number of qualified candidates needed to fill the three positions. In the example, only three names would be referred to the Chief. Now, here is that part of the decision-making process for the Officer ranks that caused the entire Affirmative Action Plan to be remanded back to the District Government for revision. The Plan required that in selecting candidates to refer to the Chief, the Promotion Board must consider the short-range goal of the Plan. The listing of selected candidates to be referred to the Chief was also to be accompanied by an explanation of how the promotions would impact upon the achievement of the short-range goals. The Plan additionally directed the Fire Chief to reject the recommendations of the Promotion Board only when the explanation given by the Board was inconsistent with the achievement of the Plan's short-range goals. For each of the Officer ranks, limits were placed on the number of times that an individual may be passed over during the two year life of the promotional lists. Finally, the Fire Department's EEO Officer of Human Rights compiles a report on the impact on the majority and female representation of each set of candidates submitted to and approved by the Fire Chief. The promotional provisions of the Plan also called for the one-time promotion of the five highest-ranking black sergeant candidates who had not been promoted from the 1982 Sergeant promotion process. It was the implementation of these five promotions that caused the International Association of Fire-fighters and Department of Justice suits.

Having described at some length the major components of the Affirmative Action Plan, I would like to present to you the District Court's findings and rationale. Briefly, the Court subjected the entry-level and promotional provisions of the plan to both Title VII and Constitutional analysis. The entry-level provisions were found to be both legal (from a Title VII standpoint) and Constitutional. The promotional provisions were found to be

in violation of Title VII. It was significant that the promotional provisions were found to meet two of the three evaluative tests used to determine the appropriateness of affirmative action processes under a Title VII framework of analysis. Because the third test was not met, the promotional provisions were found to be illegal, and the entire Affirmative Action Plan--including the entry-level provisions that passed both Title VII and Constitutional scrutiny--was disapproved and remanded back to the District Government for modification.

What were the standards and tests that were used by Judge Richey in his review of the Affirmative Action Plan? Essentially, the Court used the criteria outlined by the U.S. Supreme Court in the case United Steelworkers of America v. Weber, 443 U.S. 193 (1979) where the Supreme Court held that Title VII does not prohibit race-conscious affirmative action plans adopted by private employers. Although the Weber holding was limited to employers in the private sector, Judge Richey cited a number of cases where its Title VII analysis was extended to public employers (Vanguards of Cleveland v. City of Cleveland, 753 F. 2nd 479,484 (6th Circuit, 1985; Bushey v. New York State Civil Service Commission, 733 F. 2nd 220 (2nd Circuit, 1984), cert-denied, 109 S. Ct. 803 (1985)). For those of you who may not be familiar with the details of this seminal case, it may be worthwhile to review its facts and findings.

Prior to 1974, Kaiser Aluminum & Chemical Corporation had a plant with a skilled craftworker force of 273 persons, only 5 (1.83%) of whom were black. The local workforce, however, was 39% black. To remedy this obvious racial imbalance, Kaiser and the Union entered into a collective-bargaining agreement which contained an affirmative action plan. Kaiser established a training program to train its production workers to fill craft openings. Production workers were admitted into the training program on the basis of seniority, with the proviso that at least 50% of the new trainees were to be black. The procedures was to remain in effect until the percentage of black craftworkers in the plant approximated the percentage of blacks in the local labor force.

The plan was upheld by the Supreme Court, which gave three major reasons why the plan, although race conscious, was still permissible. First, the Kaiser plan was "designed to break down old patterns of racial segregation and hierarchy." Second, the plan did not "unnecessarily trammel" the interests of the white employees because it did not require their discharge and did not create an absolute bar to their advancement, in that 50% of the trainees would be white. Finally, the plan was temporary because it would end as soon as the percentage of blacks were in the local labor force.

Judge Richey found that those factors that made the Kaiser plan acceptable to the U.S. Supreme Court were present in the entry-level provisions of the Fire Department Affirmative Action Plan.

First, the provisions were designed to break down a pattern of racial discrimination and hierarchy. The Department was officially segregated in the past. Furthermore, an evaluation of salary levels, as well as a review of the distribution of minorities at all levels within the Fire Department, reveals underrepresentation at all levels, with the most severe underrepresentation occurring in the officer ranks. Second, Judge Richey concluded that the hiring portions of the Affirmative Action Plan constituted less of an infringement on the rights of whites than did the Kaiser Plan. Like the Weber Plan, no whites would be discharged. In addition, the plan does not prohibit the hiring of whites. In fact, with the exhaustion of each certificate, whites would be appointed in the same proportions as passed the examination. It is of great significance that Judge Richey distinguished the interests of an applicant from those of an incumbent employee. He stated that the applicants "have little expectation or entitlement to a job with the Fire Department, despite their passing the hiring examination." Third, the Judge noted with approval the temporary nature of the plan, indicating the Plan's termination date of October, 1986. As a final comment, Judge Richey gently chastised the District for failing to include a discussion of alternative entry-level certification strategies that were considered and rejected by the District. As I mentioned earlier, when the entry-level provisions of the plan were re-submitted as a separate plan for approval by Judge Richey, such a discussion of alternatives was added.

In addition to satisfying the requirements of a Title VII analysis, the entry-level provisions of the plan successfully survived constitutional scrutiny. Judge Richey found that the plan did not violate the Equal Protection Component of the Due Process Clause of the Fifth Amendment. The relevant case cited was Regents of the University of Washington v. Bakke, 438 U.S. 265,305 (1978) (opinion of Justice Powell), which indicated that a government could employ race-based classifications only when they serve a compelling governmental interest, "such as in ameliorating, or eliminating where feasible, the disabling effects of identified discrimination. Such discrimination may be identified by judicial, legislative or administrative findings..." Judge Richey cited the prior administrative finding of the entry-level examination's adverse impact on black applicants as a sufficient predicate for the implementation of race-conscious affirmative relief. In addition, Bakke was cited as recommending that the affirmative steps should "work the least harm possible to other innocent persons competing for the benefit." Finally, the case Fullilove v. Klutznick 448 U.S. 448,480 (1980) (opinion of Burger, C.J.) was cited as authority to indicate that a race conscious program designed to remedy the effects of past discrimination (must be) narrowly tailored to the achievement of that goal and that in such circumstances, "a 'sharing of the burden' by innocent parties is not impermissible." Judge Richey felt that the hirfng provisions of the Fire Department

Affirmative Action Plan were sufficiently narrowly tailored, seeking only to correct the adverse impact of the entry-level test, and to remedy past discrimination. Finally, these provisions were considered as asking the white applicants to shoulder a rather minor burden, particularly when the minimal protected interests of the applicants are considered.

As I mentioned earlier, the promotional aspects of the Affirmative Action Plan did not survive Title VII scrutiny. Judge Richey found that the provisions of the plan were temporary and justified on the basis of clear statistical evidence of a pattern of discrimination in the officer ranks. Unfortunately, he found the officer candidate referral process to be so restrictive in its focus that it effectively made race the preeminent criterion in determining those candidates to promote. Judge Richey felt that the Plan went much further than the Weber Plan, in that no whites at Kaiser were deprived of rights that they previously enjoyed, or of legitimate expectations that they had earned. Because all candidates for officer positions in the Fire Department must serve a minimum of five years before becoming eligible for promotional consideration, their protected interests were considered as much greater than those of applicants. I would like to quote directly from the opinion, now. "Any employee, in the public or private sector, who works hard and fulfills the requirements of his employment, has a legitimate expectation that he or she will be given a fair and equal opportunity to advance, based on merit and achievement. This is not something that can be taken away from him or her just because he or she happens to be of a particular race...The plan makes race a mandatory consideration over merit, and thus unnecessarily trammels the interest of white firefighters." Because the interests of white firefighters were unnecessarily trammeled, the promotional provisions, and the entire plan, did not survive its first submission to the Court. Judge Richey did note his approval in the development of a plan that would give all candidates an equal opportunity to be evaluated on their merits and ability. These concepts were kept very much in mind when the District resubmitted its promotional provisions to Judge Richey. In fact, the decision-making process was expanded to permit the Promotion Board to officially consider the candidates' performance in an oral examination that measured a variety of abilities considered critical for success as an officer. Other factors that the Board was permitted to consider included any unusual qualifications gained through experience in one or more functional areas of specialization with the Fire Department.

Finally, the short-term goal of the Affirmative Action Plan was retained but deemphasized. In fact, it was designed as one factor among many to be considered by the Board.

As public and private sector personnel practitioners, what lessons can be learned from the District of Columbia experience? First, if you wish to develop a voluntary race-conscious Affirmative Action Plan, be sure that you have properly determined that you have a problem. Be sure that you have identified the problem's cause. Build the appropriate base of documentation by engaging in utilization analysis and in adverse impact analysis. Use your administrative process. In addition to possessing compelling statistical evidence of underutilization as well as projected adverse impact, it would be helpful to have an administrative finding supporting these conclusions.

In fashioning a remedy, be sure that the action that is to be taken is curative, that it is, in fact, designed to break down a prior pattern of segregation or hierarchy. Be as careful as possible in balancing the interests of those whose interests were adversely impacted in the past with those individually innocent parties, and that it is as narrowly tailored as possible to the achievement of your goals. Be sure to limit the duration of your plan to that time when the original problem has been resolved. Be careful to include in your plan a discussion of alternative steps that were considered and rejected, along with a rationale for each rejection. Pay attention to your legal reporters. Case law and standards will continue to evolve.

These constraints may seem burdensome. Let me assure you that at the present time they will permit the development of strong, properly focused, remedial Affirmative Action Plans. Although the challenges associated with this process may appear to be substantial, I believe strongly that the results achieved are well worth the effort.

* * *

AUTHOR INDEX

Alexander, Ralph A., 67

Anderson, Claire J., 60

Baybrook, Rebecca M., 42

Boles, Stephen, 111

Boyles, Wiley R., 83

Bresler, Samuel J., 203

Broderick, Wilfrid N., 176

Brumback, Gary B., 139

Carlin, Phil A., 160

Carlisi, Anne Marie, 48

Coffee, Karen, 111

Corcione, Glenda K., 101

Davey, Bruce W., 10

Diekhoff, Foster, 71

Downey, Ronald G., 86

Dye, David A., 72,167

Ellison, Katherine W., 35

Esquilin, Jorge L., 194

Fraser, Scott L., 67

French, Jennifer, 173

Gallagher, Daniel G., 198

Giffin, Peggy, 92

Gilliland, Richard C., 134

Hambrick-Dixon, Priscilla J., 77, 192,195

Harris, Donald, 25

Hawk, John, 134

Jaffee, Cabot L., 96

Johndro, Michael W., 67

Joiner, Dennis, 111,160, 172

Juni, Esther K., 95

Lahey, Mary Anne, 80

Lang, George, 95

Love, Kevin G., 147,180

McClung, Glen G., 29

Maher, Patrick T., 33,187, 154

Mahoney, Thomas A., 12

Maurer, Robert, 120

McKillip, Richard H., 106

Maye, Doris M., 1

O'Hara, Kirk, 147

Owen, William B., 173

Palmer, Chester I., 83

Pickett, Jerry W., 129,134

Pyburn, Keith, 114

Quaintance, Marilyn K., 173

Roig, Anthony, 192

Roig, Miguel, 192

Sherman, Herbert, 39

Showers, Barbara, 5

Sproule, Charles F., 7

Stein, Eric S., 194

Steinberg, Ronnie, 53

Travers, W.T., 106

Veglahn, Peter A., 198

Veres, John G., 80

Wachter, Charles, 195

Weber, Ewald, 194